Chemical in vitro bioactivity profiles are not informative about the long-term in vivo endocrine mediated toxicity

Ingrid Greneta, Jean Paul Comet, Frédéric Schorsch, Natalia Ryan, Joerg Wichard, David Rouquié

Abstract

Endocrine disrupting chemicals raise a lot of interest and concern regarding their risk for human health and the environment. They represent a broad variety of natural and synthetic chemicals with different levels of endocrine activity evaluation. In particular, for high production volume chemicals, new methods are required to enable the evaluation of the vast number of chemicals for their potential to alter the endocrine system and prioritize them for deeper characterization. The ToxCast program from the US EPA provides data from high throughput screening assays to develop computational tools aimed at rapid in vitro bioactivity screening and prioritization.

Using publicly available data (ToxCast and ToxRef databases), we evaluate whether in vitro assay evaluations could predict in vivo outcomes observed in rat long-term studies for more than 400 chemicals. We focus on effects observed in three endocrine and two sex accessory organs and 42 in vitro assays related to pathways associated with endocrine related toxicity.

First, using simple statistical correlation we demonstrate that there is no mutual linear correlation between the selected in vitro assays and any in vivo outcome, with balanced accuracies around 50% for each assay-outcome pair. Then, by applying machine learning to investigate potential non-linear correlations, we show that the combination of different in vitro assays is not correlated with the long-term in vivo effects and cannot help to predict them since balanced accuracies are also around 50%. Moreover, the prediction based on in vitro assays is not better than the one based on classical QSAR methods. This study highlights that the selected in vitro assays do not provide information about in vivo outcomes observed in endocrine and associated organs in long-term rat in vivo studies and stresses the need for the development of in vitro assays that reflect the compounds’ pharmacokinetic properties.

1. Introduction

Since the 1990s, endocrine disrupting chemicals (EDCs) have raised a lot of interest and concern regarding their risk for human health and the environment [1]. These substances were defined by the World Health Organization as "exogenous substance or mixture that alters function(s) of the endocrine system and consequently causes adverse health effects in an intact organism, or its progeny, or (sub)populations" [2]. If sufficiently potent, these functional disruptions can lead to different adverse outcomes at the whole organism level such as developmental and reproductive effects, neurobehavioral troubles, immune disorders or cancers [3]. There are numerous and diverse mechanistic pathways that result in these effects, including activation of nuclear receptors (e.g. estrogen receptor (ER), androgen receptor (AR)),

Abbreviations: AR, androgen receptor; CART, classification and regression trees; ED, endocrine disrupting; EDC, endocrine disrupting chemicals; EDSP, Endocrine Disruptor Screening Program; EPA, Environmental Protection Agency; ER, estrogen receptor; HTS, high throughput screening; KNN, k-nearest neighbors; LDA, linear discriminant analysis (LDA); NB, Naïve Bayes; RF, Random Forest; SVM, support vector machines; SMOTE, synthetic minority over-sampling technique; BA, balanced accuracy

*Corresponding author.
E-mail address: david.rouquie@bayer.com (D. Rouquié).

https://doi.org/10.1016/j.comtox.2019.100098
alteration of steroid pathway enzymes and neurotransmitter receptors [4]. EDCs represent a broad variety of chemicals, ranging from natural, such as mycotoxins and phytoestrogens, to synthetic, such as pesticides, drugs, household products or plastics. However as of today, most of these chemicals are still lacking assessment regarding their potential endocrine activity [5]. Therefore, there is a need for new methods to enable the evaluation of the vast number of chemicals, in particular high production volume ones, for their potential to alter the endocrine system and prioritize them for deeper characterization.

For this purpose, the US Environmental Protection Agency (EPA) created the Endocrine Disruptor Screening Program (EDSP) in 1999 in which they proposed to test chemicals in a battery of five in vitro and six in vivo Tier 1 tests (EPA EDSP link – https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-tier-1-battery-assays). The in vitro tests focus on ER, AR and steroidogenesis pathways. The first in vivo ED screens include the uterotrophic assay (UTA), the Hershberger assay and male and female pubertal assays. Moreover, in vivo multigenerational Tier 2 tests were also proposed to further characterize compounds that were active in the Tier 1 tests. However, since the battery of assays proposed in the EDSP program is animal consuming, they are low-throughput and do not allow the evaluation of potential endocrine disruption (ED) effects of a large number of compounds.

In 2007, the ToxCast program was established by the US EPA in order to generate and use data from high-throughput screening (HTS) assays and develop computational tools. The main goal of this project is to rapidly screen compounds for in vitro bioactivity and prioritize for further testing. The targets of the HTS assays are diverse and cover a large number of biological pathways, including ER, AR and steroidogenesis pathways. In particular, high-throughput alternatives to parts of the EDSP testing battery are available including the ER and AR binding assays, the ER transactivation assay, the aromatase assay and the steroidogenesis assay performed in H295R cells.

In total, 18 and 12 in vitro assays have been implemented within the ToxCast program to target key events along the ER and AR pathways, respectively. Researchers have integrated these assays into two computational linear additive models in order to discriminate between chemicals that are true agonists or antagonists of the pathways and the ones that are false positives because of assay-interference and cytotoxicity [6–8]. They compared the results obtained from their models to existing in vitro and short-term in vivo public data for the UTA and Hershberger studies and obtained 84–93% accuracy for the ER model and 95–97% for the AR model. Thus, they were able to validate both models and conclude that there is a high correlation between the predictions based on ToxCast assays and the short-term in vivo effects observed in estrogen and androgen dependent tissues.

To go a step further, we decided to assess the potential of compounds to act as EDCs in long-term in vivo studies, based on the in vitro ToxCast assay data. Among the in vitro data publicly available, the Toxicity Reference Database (ToxRefDB) provided by the EPA captures results from thousands of in vivo toxicity studies performed in laboratory animals for hundreds of compounds [9].

In the last decades, in silico models have been developed to link compounds’ structure to in vitro activity [10] but the relationship between in vitro results and in vivo outcomes has not been fully explored, especially for specific toxic outcomes such as ED.

In this work we study the link between in vitro bioactivity from ToxCast assays and adverse outcomes observed in rat long term studies, obtained from ToxRefDB. We focus on in vitro assays related to ER (E), AR (A) and steroidogenesis (S) pathways and in vivo effects observed in adrenal glands and reproductive tract tissues (testes, ovaries, prostate and uterus). First, we look at the correlation between each in vitro assay and the in vivo outcomes and compare it with the correlation between the published ER and AR model results and the same in vivo outcomes. This analysis demonstrates that there is no mutual linear correlation between the in vitro assays and any in vivo outcome. Then, as an extension of the first order correlation, we built machine learning (ML) models to predict the in vivo outcomes, either based on the in vitro assays alone, the chemical structure alone or a combination of both. Our results highlight that, based on the results of more than 400 compounds, ToxCast in vitro assays that are related to pathways altering endocrine activity do not discriminate compounds which actually lead to long-term in vivo toxicity in the selected endocrine-related organs.

2. Materials and methods

2.1. Data sources

Three types of datasets were used in this study: an in vitro dataset, an in vivo dataset and the chemical structure of the compounds. All data are publicly available and were accessed through the EPA website: https://www.epa.gov/chemical-research/toxicity-forecasting.

2.1.1. In vivo toxicity data

In vivo data were obtained from the Toxicity Reference Database (ToxRef DB) released in October 2014 which gathers data from in vivo toxicological studies performed on hundreds of compounds in several species of laboratory animals and for different time periods. Results are provided as the NOAEL or LOAEL (No/Lowest Observed Adverse Effect Level) for each type of toxic effect reported.

In our study we focused on outcomes observed in rat long term studies, referred to as “CHR” studies in ToxRefDB. Of the CHR studies, 80% are 2-year rat carcinogenicity studies and 20% are 13-week to 31-month studies. We only used studies that have been referred to as “acceptable guideline” in the database (i.e. studies are complete and meet official guideline requirements), therefore retaining studies for 445 compounds. We focused on outcomes observed in five endocrine related organs: adrenal glands, ovary, testis, prostate and uterus. For each organ, we classified and grouped the observed effects listed in ToxRefDB into adverse effect categories: three categories for adrenal glands (steroidogenesis effects, stimulation, injury), two for ovary (effect on germinal cells, effect on interstitial cells), two for testis (effect on germinal cells, effect on spermatogenesis), one for prostate and one for uterus. For each category independently, we used these in vivo results as binary data by assigning 1 to compounds that had an effect (whatever the corresponding dose) and 0 otherwise.

2.1.2. In vitro bioactivity data

In vitro data were obtained from the ToxCast database (October 2015 release) which gathers the results of 1192 HTS assays performed on 9076 compounds during ToxCast Phase I and II. Results are provided as the EC50 which is the concentration (in micromolar) corresponding to the half maximal efficacy. Of the 445 compounds selected from ToxRefDB, 418 were found in the ToxCast database.

As the current study focuses on pathways leading to endocrine disruption, we selected in vitro assays related to AR, ER, aromatase, steroidogenesis and other receptors located in endocrine organs. We manually performed this selection, based on expertise and knowledge of endocrine pathways. For ER and AR, we referred to the assays used by the published computational models [6,8]. In an effort to ensure a robust dataset with enough representatives of active versus inactive compounds, we applied a filter to keep only assays that have at least five percent of active compounds (for the 418 compounds overlapping between ToxCast and ToxRefDB) and ended up with 42 assays. Table 1 provides the list of assays, the associated pathway and the assay type. In summary we used 12 assays related to ER, 9 related to AR, 2 related to aromatase, 11 related to steroidogenesis and 8 related to other receptors.

2.1.3. Chemical structure

Chemical structure information for 8599 unique substances was obtained from DSSTox (Distributed Structure – Searchable Toxicity)
Table 1

<table>
<thead>
<tr>
<th>Assay name</th>
<th>Pathway</th>
<th>Type of assay</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACEA_T47D_80hr_Positive</td>
<td>E</td>
<td>Cell proliferation</td>
</tr>
<tr>
<td>ATG_ER,E,G,CIS_up</td>
<td>E</td>
<td>mRNA induction</td>
</tr>
<tr>
<td>ATG_ER,TRANSEP_up</td>
<td>E</td>
<td>mRNA induction</td>
</tr>
<tr>
<td>OT_ER,ERaERb_0480</td>
<td>E</td>
<td>Protein complementation</td>
</tr>
<tr>
<td>OT_ER,ERaERb_1440</td>
<td>E</td>
<td>Protein complementation</td>
</tr>
<tr>
<td>OT_ER,ERaERb_1440</td>
<td>E</td>
<td>Protein complementation</td>
</tr>
<tr>
<td>OT_ES,ERaERb_0480</td>
<td>E</td>
<td>Protein complementation</td>
</tr>
<tr>
<td>OT_ES,ERaERb_1440</td>
<td>E</td>
<td>Protein complementation</td>
</tr>
<tr>
<td>OT_Era,ERGFP_0120</td>
<td>E</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>OT_ErE,ERGFP_0480</td>
<td>E</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>TOX21,Era,BLA,Antagonist_ratio</td>
<td>E</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>TOX21,Era,LUC,BG1,Agonist</td>
<td>E</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>TOX21,Era,LUC,BG1,Antagonist</td>
<td>E</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>NVS_NR,hAR</td>
<td>A</td>
<td>Receptor binding</td>
</tr>
<tr>
<td>NVS_NR,rAR</td>
<td>A</td>
<td>Receptor binding</td>
</tr>
<tr>
<td>OT_AR,ARELUC,AG,1440</td>
<td>A</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>OT_AR,AR,5C1_0480</td>
<td>A</td>
<td>Coregulator recruitment</td>
</tr>
<tr>
<td>TOX21_AR,BLA,Antagonist_ratio</td>
<td>A</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>TOX21_AR,MDA,B2,Antagonist</td>
<td>A</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>TOX21_AR,MDA,B2,Antagonist2</td>
<td>A</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>CEETOX,HO295,11DCORT,dn</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,ANDR,dn</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,CORTISOL,dn</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,DOC,dn</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,ESTRA101L,dn</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,ESTRONE,dn</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,ESTRONE,up</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,OHPROG,dn</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,OHPROG,up</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,PROG,up</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>CEETOX,HO295,TESTO,dn</td>
<td>S</td>
<td>Hormone measurement</td>
</tr>
<tr>
<td>NVS_ADME,HCYP19A1</td>
<td>S</td>
<td>Enzyme activity</td>
</tr>
<tr>
<td>TOX21,Aromatase_Inhibition</td>
<td>S</td>
<td>Enzyme inhibition</td>
</tr>
<tr>
<td>ATG_Sp1,CIS_up</td>
<td>O</td>
<td>mRNA induction</td>
</tr>
<tr>
<td>ATG_GRE,CIS,dn</td>
<td>O</td>
<td>mRNA induction</td>
</tr>
<tr>
<td>ATG_SREBP,CIS_up</td>
<td>O</td>
<td>mRNA induction</td>
</tr>
<tr>
<td>NVS_NR,hBG</td>
<td>O</td>
<td>Receptor binding</td>
</tr>
<tr>
<td>NVS_NR,hGR</td>
<td>O</td>
<td>Receptor binding</td>
</tr>
<tr>
<td>NVS_NR,hPR</td>
<td>O</td>
<td>Receptor binding</td>
</tr>
<tr>
<td>TOX21_GR,BLA,Agonist_ratio</td>
<td>O</td>
<td>Reporter gene</td>
</tr>
<tr>
<td>TOX21_GR,BLA,Antagonist_ratio</td>
<td>O</td>
<td>Reporter gene</td>
</tr>
</tbody>
</table>

SDF (Structure Data File) files (October 2014 release). These files contain compound structure, name, CASRN, SMILES, and other information that was not needed for the current analysis.

2.2. Statistical analysis – correlation

2.2.1. Correlation analysis

In vitro assay results (AC50, as reported in the US EPA ToxCast database [11]) were turned into binary values: a value of 1 was used if the ToxCast data analysis pipeline determined the chemical-assay pair to be active, and 0 otherwise.

For each pair of in vitro assay and in vivo outcome, we computed the three following metrics:

- Sensitivity = TP/(TP + FN)
- Specificity = TN/(TN + FP)
- Balanced Accuracy (BA) = (Sensitivity + Specificity)/2

where TP (respectively TN) is the number of True Positive (respectively True Negative) compounds, i.e. compounds which are positive (respectively negative) for both in vitro assay and in vivo outcome; and FP (respectively FN) is the number of False Positive (respectively False Negative) compounds, i.e. compounds which are positive (respectively negative) in vitro but negative (respectively positive) in vivo.

All statistical analysis was performed using R version 3.2.3 software.

2.2.2. Results from ER and AR computational models

In order to determine if aggregating several in vitro assays is more predictive of in vivo outcomes than a single in vitro assay, we used the results from EPA’s computational models for ER [6] and AR activity [8].

Basically, these models sum the activity of a chemical obtained for each essay that contributes to the model in a non-weighted manner and for different concentrations. Thereby, for each chemical and concentration, a linear sum of the activity measured in the 18 (respectively 12) in vitro assays targeting the ER (respectively AR) pathway is computed. In the end, the activity is described by a concentration response curve for which the area under the curve (AUC) is measured. The AUC ranges from 0 to 1 and is referred to as the score of the model. Detailed methods and results, including the AUC score, quality criteria and flags, are available for both ER and AR models [6–8].

For the ER model, two scores are computed to assess the quality or reliability of the AUC score. A Z-score flags non selective assay activity due to cytotoxicity by measuring the distance between the AC50 obtained for a compound in an assay of the ER model and the AC50 obtained for this chemical in the cytotoxicity assays, and factors in the variability across all chemicals and all cytotoxicity assays. A low distance (Z-score < 3) indicates that the activity measured in the considered assay could be due to cytotoxicity and not to a target-selective mechanism. The T-score corresponds to the maximum activity measured (i.e. the highest point of the concentration-response curve). Indeed, since concentration-response curves are normalized compared to a control or baseline, the maximal activity is a relative percentage that not necessarily equal to 100%. The T-score therefore corresponds to the highest value of this relative percentage. For both Z and T scores, a median was computed across all ER assays for each compound and referred to respectively as med.Z and med.T [6].

According the AR model, a confidence score is provided which takes into account the AUC value of the model, the same Z-score as for the ER model and the results of a supplemental assay which can confirm the antagonist activity of chemicals.

Among the 418 compounds from our study, 361 have a score available for the ER and AR models.

To discriminate between positive and negative compounds for these models, we chose the following thresholds for the different values available:

- Positive for ER model if model score > 0.1 (either agonist or antagonist activity) and med.T > 50% and med.Z > 3; negative otherwise
- Positive for AR model if model score > 0.1 (either agonist or antagonist activity) and, in the case of an antagonist activity, confidence score > 0 (this confidence score is not provided for the agonist activity, therefore all the agonists are considered as positive when model score > 0.1); negative otherwise

In the end, 5 compounds were positive for the ER model and 55 for the AR model.

We performed the correlation analysis for the ER and AR models as described above for the 42 assays independently and plotted the results on the same graphs as a comparison. Note that from the 19 assays used by Judson et al. in their model for ER, 7 are excluded from our study because their hit rate was below our cutoff of 5%. This is also the case for 3 assays of the 12 used by Kleinstreuer et al. for the AR model.

2.3. Machine learning

We used machine learning methods to predict the in vivo outcomes observed in endocrine organs from either the structure of compounds and/or their in vitro bioactivity.
The machine learning approach was based on the work of Liu et al. [11] and the associated published Python code.

2.3.1. Datasets
Since the machine learning methods used in this approach are not good at handling missing data we identified a complete matrix between all in vitro and in vivo data (i.e., all the compounds in the dataset have been tested in all the 42 assays previously selected). From the 418 compounds available, 341 compounds met this criteria.

2.3.2. Chemical structure descriptors
The Structure Data Files (SDF) for the 341 compounds were obtained from the EPA website. After cleaning the structures of the compounds (removing salts and inorganic elements, neutralizing and checking for duplicates), we computed two types of chemical structure descriptors:

- 74 molecular descriptors using RDKit which are physico-chemical properties. Continuous values were normalized between 0 and 1;
- 731 fingerprints using pybel package in Python [12] and PaDEL software [13]: FP3 (18), Estate (27), KlekotaRoth (184), PubChem (331), SubFP (43), MACCS (128).

2.3.3. Bioactivity descriptors
The 42 in vitro assays as selected were used as individual descriptors of the bioactivity of compounds. As in Liu et al., we set $AC_{50}$ values of inactive compounds to $1 \times 10^6 \mu M$ and transformed all the $AC_{50}$ according to the following formula: $AC_{50} = 6 - \log_{10}(AC_{50})$. This formula gives inactive compounds a value of 0 and represents active ones on a continuous ascending scale and reflects their potency. Then the values were normalized between 0 and 1.

2.3.4. Machine learning
The machine learning method was implemented in Python2.7 and described by Liu et al. [11].

We used the same five classification algorithms as Liu (linear discriminant analysis (LDA), Naïve Bayes (NB), support vector machines (SVM) with two different kernels (linear and radial basis function), classification and regression trees (CART), k-nearest neighbors (KNN)) with the same default parameters. We also used the Random Forest (RF) algorithm (with default parameters and 100 trees) because it is an ensemble method that performs better in terms of generalization than a single regression tree, can handle many input features and lowers risk of overfitting [14]. Finally, we used an ensemble technique (ENSMB) [15] for which the prediction corresponds to the majority vote of the six previous classifiers.

A 10-fold cross-validation testing was performed and repeated 20 times. For each step in the cross-validation, the descriptors were ranked by computing their importance score using the Random Forest attribute feature importance (this was different from Liu et al. who computed the univariate association between each pair of descriptors and the in vitro outcome). Then, classifiers were built using the 10–42 or 60 best descriptors, (depending on the type of descriptor: 42 when only in vitro assays are used and 60 when chemical descriptors are used), by adding one descriptor at each step.

The results for each category reported below are for the model that showed the highest BA with its corresponding number of descriptors used.

2.3.5. Data augmentation
Since the datasets in this study are imbalanced (more inactive compounds than active ones) we used data augmentation to rebalance the data and build new classifiers. We utilized the technique SMOTE (Synthetic minority over-sampling technique) [16] which aims at creating new synthetic samples based on linear interpolation of actual data. Basically, for each observation (i) of the minority class, it randomly selects one of its k-nearest neighbors (j) of the minority class in the descriptors space and generates a random example that is along the line between $i$ and $j$ according to the following formula: $x_{new} = x_i + (x_j - x_i)\delta$ where $x$ corresponds to the vector of descriptors of the different observations and $\delta$ is a random number from the interval [0,1]. This process is repeated for all or part of the k-nearest neighbors of each observation from the minority class, according to the desired final number of new samples.

We used this technique in each step of the cross-validation loop in order to increase the number of compounds of the minority class of the training set.

2.3.6. Performance evaluation
The performance of the classifiers was evaluated using the three metrics described above: sensitivity, specificity, and balanced accuracy (BA).

3. Results
3.1. Overview
The overall goal of the study was to assess if the evaluation of chemicals in in vitro assays could provide information about in vivo endocrine-related effects observed in long-term carcinogenicity rat toxicity studies (Fig. 1). For this we used the publicly available datasets ToxCast and ToxRefDB. We focused on effects observed in three endocrine organs (adrenal glands, testis and ovary) and two sex accessory organs (prostate and uterus). From the available in vitro assays in ToxCast, we identified 42 assays that should be most informative and predictive of endocrine effects based on their relation to the following biological pathways: estrogen receptor (E), androgen receptor (A), steroidogenesis pathway (S) or other endocrine related receptors (O).

First, we looked at the correlation between each pair of in vitro assays and in vivo effects observed in the selected target organs. Second, in an attempt to predict in vivo effects we applied machine learning using in vitro assay results, chemical structure information or a combination of the two.

3.2. Statistical analysis
3.2.1. Data
In total, 418 compounds have in vivo data for chronic rat studies in ToxRefDB and have been tested in ToxCast, but not always in all the 42 selected in vitro assays. For each of the 42 assays, Table 2 provides the total number of compounds tested with the number of positive and negative ones as well as the corresponding percentage of actives.

From the 418 compounds, 349 have been tested in all the 42 selected in vitro assays. The results show a range of percentage of active only between 5 and 30% (mean 12.7%) indicative of a highly imbalanced dataset in favor of negative compounds.

Regarding the computational models for ER and AR (which aggregate several in vitro assay results into one model that generates an ER and AR score), scores are available for 361 compounds among the 418. We applied filters to discriminate between positives and negatives (see Materials and Methods) and only 5 compounds were positive among the 361 for the ER model (1.4%) and 55 for the AR model (15%).

Table 3 summarizes the in vivo data used with the number of positive and negative compounds for each of the 9 effect categories for the 5 organs. Here again the data are highly imbalanced in favor of negative compounds with 4 to 16% of positive compounds depending on the category.
3.2.2. Statistical analysis results

For each in vivo outcome, we plotted the sensitivity, specificity and balanced accuracy computed to look at the correlation with all the 42 in vitro assays and the ER and AR models (Fig. 2).

We observed that for all pairs of assay and in vivo outcome, the specificity was high (between 0.85 and 0.95) and the sensitivity was very low (lower than 0.3) leading to an overall BA around 0.5. Of note, ER pathway related assays (E) did not show a higher BA for ovary and uterus outcomes compared to the other assays. The same was observed for the AR and steroidogenesis pathways related assays (A and S) and testis, prostate and adrenal gland outcomes, respectively.

When looking at the correlation between the computational model for AR [8] and in vivo outcomes (in particular in prostate and testis), we did not see any difference compared to the individual AR pathway related assays. This result showed that even the aggregation of several assays related to the AR pathway does not improve the correlation with the selected in vivo outcomes that were examined here. Regarding the computational model for ER, we did observe an increase of BA and sensitivity for 4 outcomes (Uterus, Spermatogenesis testis, Germinal cells ovary and Prostate) but since there is only 1.4% of positive compounds for this model in our datasets, it is difficult to know the significance of this finding.

Overall, this simple statistical analysis demonstrated that there is no mutual linear correlation between the 42 in vitro assays and any of the selected in vivo outcomes when we used the results of each assay independently. Somewhat surprisingly, this observation was also made for the in vitro assays with targets physiologically related to specific in vivo adverse outcomes (e.g. in vitro AR assays were not correlated with effects observed in prostate or testis known to result from a perturbation of the AR pathway). Moreover, the use of a linear additive model to consider several assays in the same biological pathway as proposed by Judson et al. and Kleinreuter et al. [6,8] did not show a higher correlation for the selected long-term adverse outcomes, but we could only draw a clear conclusion for the AR model.

We then investigated if there was a possibility to predict in vivo outcomes based on a combination of several of the 42 in vitro assays using machine learning methods.

3.3. Prediction of in vivo toxicity outcomes based on in vitro assays or/and structural descriptors using machine learning methods

Machine learning models were built to predict the 9 effect categories for the 5 organs (Table 3). For each category of effects, three types of models were built depending on the input descriptors: biological (the 42 in vitro assays), chemical structure (physico-chemical properties and fingerprints) or a combination of both. Since the data were imbalanced, we also used a data augmentation technique to test if it could improve the predictive model performance.

3.3.1. Datasets

As described above, 341 compounds had results in ToxRefDB for long term rat studies and have been tested in all the 42 ToxCast assays. Table 4 summarizes the number of positive and negative compounds in the datasets used for each in vivo effect category for the 5 organs. As already described by Liu et al., we chose to be quite stringent and call a compound negative (assigned a value of 0) for a specific organ only if it was negative for all the organ’s effects categories. For example, if a compound did not induce any of the 3 category effects in the adrenal glands (Steroidogenesis effects, Stimulation or Injury), it was considered “negative”. However, if a compound induced one of these 3 category effects (e.g., Stimulation), it was considered positive in the “Stimulation” dataset but discarded from the two other adrenal glands datasets (Steroidogenesis effects and Injury). Therefore, not all the 341 compounds are represented in each dataset.

For all the endpoints considered, the datasets for machine learning were highly imbalanced in favor of negative compounds. The lowest
The percentage of actives was for the effects in the prostate (4.5%) and the highest was for stimulation in adrenal glands (19%).

3.3.2. Results

Figs. 3–5 show the performance of the models obtained for the 9 effects with the Random Forest (RF) algorithm, before and after SMOTE (data augmentation technique to reduce the impact of the unbalanced nature of the dataset used). We chose to present the results of RF because of its ability to handle numerous features and it is commonly accepted as an algorithm avoiding overfitting. The results obtained with all the other methods are available in Supplemental files.

3.3.2.1. Machine learning based on in vitro assays alone. For the adrenal outcomes (Fig. 2), ovary and uterus outcomes (Fig. 3) and testis and prostate outcomes (Fig. 4), all had low sensitivity and high specificity, and BA lower than 0.6. The SMOTE method increased sensitivity but lowered specificity leading to same overall BA.

For all effects, with and without using the SMOTE method, we observed that the sensitivity is very low (between 0.05 and 0.09) and the specificity is high (between 0.95 and 0.99) which led to a BA between...
Fig. 2. Results of statistical analysis. Balanced accuracy (pink), sensitivity (blue) and specificity (purple) between each of the 42 in vitro assays or one of the two EPA computational models (ER and AR) and the in vivo outcomes observed after rat chronic studies in adrenals, ovaries, uterus, testis and prostate. E: estrogen pathway related assays (including ER model), A: androgen pathway related assays (including AR model), S: steroidogenesis pathway related assays, O: other assays.

(a) Steroidogenesis adrenals
(b) Stimulation adrenals
(c) Injuries adrenals
(d) Effects on germinal cells ovary
(e) Effects on interstitial cells ovary
(f) Effects on uterus
When using the SMOTE method, the sensitivity was increased (between 0.1 and 0.49) but the specificity was decreased (between 0.58 and 0.94), resulting in a BA still around 0.50. Similar results were obtained with the other algorithms (see Supplemental data).

Overall, these results showed that machine learning models to predict in vivo effects observed in endocrine organs from the selected in vitro assays did not perform better than chance (BA around 0.5) and that data augmentation did not help to increase the performance. This highlighted that a combination (linear or not) of different in vitro assays is also not correlated to the selected long-term in vivo effects and cannot help to predict them.

### 3.3.2.2. Machine learning based on chemical structure information alone or combined with in vitro assays

Regarding the prediction of the in vivo outcomes from either the chemical structure information alone or combined with in vitro assays. Regarding the prediction of the in vivo outcomes from either the chemical structure information alone or combined with in vitro assays. Regarding the prediction of the in vivo outcomes from either the chemical structure information alone or combined with in vitro assays. Regarding the prediction of the in vivo outcomes from either the chemical structure information alone or combined with in vitro assays.

#### Table 4

<table>
<thead>
<tr>
<th>Organ name</th>
<th>Endpoint</th>
<th>Number of positive compounds (percentage)</th>
<th>Number of negative compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uterus</td>
<td>Uterus effect</td>
<td>33 (9.7%)</td>
<td>308</td>
</tr>
<tr>
<td>Prostate</td>
<td>Prostate effect</td>
<td>15 (4.5%)</td>
<td>326</td>
</tr>
<tr>
<td>Adrenal glands</td>
<td>Steroidogenesis effects</td>
<td>51 (16%)</td>
<td>264</td>
</tr>
<tr>
<td></td>
<td>Stimulation</td>
<td>62 (19%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Injury</td>
<td>47 (15%)</td>
<td></td>
</tr>
<tr>
<td>Ovary</td>
<td>Effect on germinal cells</td>
<td>25 (7.5%)</td>
<td>307</td>
</tr>
<tr>
<td></td>
<td>Effect on interstitial cells</td>
<td>29 (8.6%)</td>
<td></td>
</tr>
<tr>
<td>Testis</td>
<td>Effect on germinal cells</td>
<td>56 (17%)</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td>Effect on spermatogenesis</td>
<td>33 (11%)</td>
<td></td>
</tr>
</tbody>
</table>
combined with the \textit{in vitro} assays, we observed that the performances of the models are similar to the ones of the models built with \textit{in vitro} assays alone (Figs. 3, 4 and 5). Regarding the impact of the data augmentation technique, the sensitivity was not increased and the specificity was not decreased as it was for the model built on \textit{in vitro} assays alone. For example for the outcome “Steroidogenesis in adrenal glands”, sensitivity and specificity were around 0.1 and 0.99 before data augmentation respectively, and around 0.1 and 0.95 after data augmentation, either with chemical structure alone or combined with \textit{in vitro} assays. However, since the shown plots provide the performance only for the model which provided the best BA but not the best sensitivity, we may observe higher sensitivity for other models but the BA would be lower (e.g. the best sensitivity obtained for steroidogenesis in adrenal glands is 0.15 and corresponding BA of 0.51).

Overall, these results showed that using chemical structure to predict the long-term \textit{in vivo} effects is neither better nor worse than using...
4. Discussion/conclusion

In this work we evaluated the ability of in vitro assay results (ToxCast program) to predict in vivo outcomes observed in rat long-term studies (ToxRef database). Our analysis utilized 404 chemicals, 42 in vitro assays related to endocrine pathways and in vivo endpoints from three endocrine organs (adrenal glands, ovary and testis) and two sex accessory organs (uterus and prostate).

Using simple statistical linear correlation and machine learning methods in order to investigate potential non-linear correlations, we were able to show three main conclusions. First, the 42 selected in vitro assays were not correlated to the in vivo outcomes, even for assays that were specific for relevant pathways known to be present in the target organs. Indeed, ER related assays were not more correlated with outcomes observed in ovary or uterus than in the other organs. The same was observed for AR related assays and prostate and testis in vivo outcomes observed in ovary or uterus than in the other organs. The same was observed for AR related assays and prostate and testis in vivo outcomes observed in ovary or uterus than in the other organs.
outcomes and for steroidogenesis related assays and adrenal glands outcomes. To determine if better results could be obtained by considering results from multiple assays, we also performed the analysis using the results from the published computational models for ER and AR pathways that aggregate several assays into one single score for each pathway. For the AR model, we could conclude that the use of a linear additive approach that considers several assays in the same pathway did not show a higher correlation with long-term in vivo outcomes compared to predictions made using one single assay. Regarding the ER model, we could not draw a valid conclusion because of the small number of positive compounds (1.4%) in our analysis.

Second, machine learning models built on the data from the 42 in vitro assays were not able to predict the in vivo effects. One significant limitation to building accurate models was the unbalanced nature of the datasets used in this work (hit rates < 10%). A data augmentation technique was applied but it did not improve the predictive ability of the models. Together, these results suggest that the combination of different in vitro assays is not correlated to the long-term in vivo effects.

![Fig. 5. Performance of ML models that predict testis and prostate outcomes using RF algorithm and that reached the highest BA for each type of descriptors.](image)
and cannot help to predict them, even if the combination is not linear.

Finally, the predictions from machine learning models built on the selected 42 \textit{in vitro} assays are not better than those derived from chemical structure alone for \textit{in vivo} effects. Furthermore, a combination of both types of descriptors also did not improve the performance. This leads us to conclude that these \textit{in vitro} assays do not provide information about \textit{in vivo} outcomes observed in endocrine and associated organs in rat long-term studies.

Although it may initially be discouraging to find that \textit{in vitro} assays are currently unable to predict long-term endocrine outcomes \textit{in vivo}, there are several factors (both related to the data and to biology) that help explain the results of our study, and suggest areas for further development of this type of research. First, our analysis was based on a relatively small number of compounds (418). The data used were limited to the publicly available \textit{in vivo} rat carcinogenicity studies. This illustrates a general challenge faced by the computational toxicologist regarding the small volume of data publicly available and the importance of initiatives such as eFOX that aim at gathering, organizing and making available \textit{in vivo} toxicological data [17]. Further, toxicological data, from both \textit{in vitro} and \textit{in vivo} models, is often highly imbalanced and, therefore, not well suited for computational purposes and machine learning in particular. This small number of compounds combined with the imbalanced property of the data also led to the removal of some ER and AR related \textit{in vitro} assays when applying the cutoff of a minimum of 5% of positive compounds. Thus, we do not know if these assays could have contributed to a better prediction of the long-term \textit{in vivo} effects. Another issue faced when utilizing \textit{in vivo} toxicological data sources is a general lack of well harmonized ontology, which has been fully discussed by others [18,19]. The same endpoint or finding could be referred to by several different terms depending on the laboratory or pathologist that conducted the study. We made our best attempt to address this issue by grouping the different endpoints into categories to make the effects more inclusive, but we expect that our methods would have had better performance with more precise and harmonized histopathological ontology. Various organizations are making progress in developing terminology and ontologies (e.g., the US Food and Drug Administration (FDA) utilizing Standard for Exchange of Nonclinical Data (SEND) [18]; the Society of Toxicology Pathology and the European Society of Toxicology leading the International Harmonization of Nomenclature and Diagnostic (INHAND) criteria for lesions in rats and mice project [20]), but global harmonization has not yet been achieved. Finally, databases of \textit{in vivo} toxicology data may only include effects used to determine the lowest observed adverse effect level (LOAEL), meaning that additional effects at higher doses are not reported and not available for data analysis exercises. In particular, this has been pointed out for reproductive toxicity studies in ToxRefDB [20].

In addition to characteristics of the data utilized in the analysis, the biological meaning of the datasets also contributes to the outcome of our study. Several aspects can explain why a compound is called active \textit{in vitro} and negative \textit{in vivo} and conversely. First, it is important to state that the \textit{in vitro} assays used here do not give information about ADME (Absorption, Distribution, Metabolism and Excretion) properties and therefore the results do not reflect dose dependencies in the \textit{in vivo} context. These properties are critical to obtain accurate \textit{in vivo} predictions, as recently highlighted [21,22]. Also, the selected 42 assays do not represent the set of all possible biological pathways leading to adverse endocrine effects. Assays selected in the ToxCast project were not originally designed to be predictive of specific long-term \textit{in vivo} outcomes or toxicological modes of action [23]. Furthermore, intercellular and inter-organ communication is also not captured by \textit{in vitro} assays [24] which prevents detecting the \textit{in vivo} responses that require multi-tissue interactions [25]. This is often crucial in endocrine mediated toxicity given the key compensatory role of the pituitary gland for many endocrine related tissues [26].

In the last years, most of the computational work performed to predict toxicological effects and bioactivity of compounds has tried to link compound structures to \textit{in vitro} or \textit{in vivo} data by applying read across approaches [27] or machine learning methods [28,29] and recently a combination of both [30]. Nonetheless, the link between \textit{in vitro} and \textit{in vivo} data has not been evaluated much and was considered either broadly [11,31] or for specific outcomes [32,33]. In particular, Thomas et al. also used ToxCast and ToxRefDB to evaluate how \textit{in vitro} assays and/or chemical structure could classify 60 \textit{in vivo} toxicity endpoints using statistical and machine learning approaches [34]. They were not able to obtain accurate predictions and showed that machine learning models based on \textit{in vitro} assays alone did not perform better than those based on structural descriptors and that the combination of both types of descriptors did not improve the performance. The ToxCast team replied to this work and suggested that biological knowledge should be used to build such models [35]. This is what we tried to do by performing a pre-selection of \textit{in vitro} assays that could be related to specific \textit{in vivo} outcomes. Moreover, recent studies showed that \textit{in vitro} bioactivity data was not able to correctly predict carcinogenicity in rodents or cancer hazard classification [36,37]. Here we proposed for the first time to look at the link between specific \textit{in vitro} assays and \textit{in vivo} effects arising in rat endocrine related organs after long-term exposure using simple correlation and machine learning methods. This work extends the evaluation of EPA’s additive models for ER and AR pathways, aiming at the replacement of EDSP Tier I assays which includes short term \textit{in vivo} pubertal assays, uterotrophic assay and Hershberger assay [5,6,7], by evaluating if \textit{in vitro} assays could give information about the \textit{in vivo} long-term ED effects. In particular, since the EPA’s additive models resulted in good predictions of uterotrophic and Hershberger assay, our results suggest that, in extension, these short-term \textit{in vitro} assays are probably not good to predict long-term endocrine-related \textit{in vivo} endpoints such the ones we are interested in here. It would have been interesting to look at the number of true positive and true negative compounds between the two short-term assays and the long-term endocrine-related effects. Unfortunately, it was not possible as too few compounds were found having both long-term and Uterotrophic and Hershberger data available. Finally, it is worth to mention that in essence, an \textit{in vivo} evaluation has got its own level of uncertainty regarding the reproducibility as it has already been described [39,40].

In general, since the results of our work showed that there is no evident link between the 42 \textit{in vitro} assays selected and \textit{in vivo} effects observed in rat long term studies, we would suggest being cautious when interpreting the meaning and relevance of positive \textit{in vitro} assays. Ideally, a hypothesis-driven approach should be conducted to drive the selection of appropriate \textit{in vitro} assays specifically addressing the prediction of a given adverse outcome pathway in order to ensure the causal link between the two types of evaluations [41]. Further, including PBPK modelling in this approach could allow for the consideration of dose in the predictions from \textit{in vitro} to \textit{in vivo}. Indeed, physiologically based kinetic modelling-based reverse dosimetry has recently been shown to accurately simulate the dose-response of \textit{in vivo} uterus growth induced by estrogenic compounds using \textit{in vitro} data as input information [38].

Moreover, our results highlighted the important need for more publicly available data, with harmonized results and ontology, to be utilized in computational and predictive toxicology efforts.

In conclusion, statistical analysis and machine learning based on the results of more than 400 compounds showed that ToxCast \textit{in vitro} assays that are related to pathways altering endocrine activity do not discriminate compounds which actually lead to long-term \textit{in vivo} toxicity in endocrine organs in the reproductive tract (testis, prostate, ovary, uterus) as well as in the adrenal glands.

Future work should address the lack of description of toxicokinetic properties of compounds in order to enable \textit{in vitro-to-in vivo} extrapolation (IVIVE) and estimate the dose of effective potency as already suggested by Thomas et al. [22]. Also, more efforts are required to
develop specific in vitro assays informing about a broader spectrum of pathways leading to endocrine-mediated adverse outcomes, in particular in the case of toxic modes of action involving several organs [26,42]. Furthermore, this type of computational study predicting in vivo effects from in vitro data should be expanded to other types of adverse effects observed as well as to other laboratory animal species.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgment

I.G. holds a doctoral fellowship from the Association Nationale de la Recherche Technique (ANRT CIFRE PhD funding).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.comptox.2019.100098.

References


