iPRIOR: Online Public Platform for Modeling ToxCast™ Assays and Prioritization of Animal Toxicity Testing

Ahmed Abdelaziz[1§], Yurii Sushko[1], Sergii Novotarskyi[1], Robert Körner[1], Stefan Brandmaier[2], Igor V. Tetko[1,3]

[1]eADMET GmbH, Lichtenbergstraße 8, D-85748 Garching / Munich, Germany; [2]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany;
[3]Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, Munich, Germany;

[§]Corresponding author

Telephone: +4915776887277
Fax: +4989716801649
Email address: contact@amaziz.com

ABSTRACT: The use of long-term animal studies for human and environmental toxicity estimation is more discouraged than ever before. Alternative models for toxicity prediction, including QSAR studies, are gaining more ground. A recent approach is to combine in vitro chemical profiling and in silico chemical descriptors with the knowledge about toxicity pathways to derive a unique signature for toxicity endpoints. In this study we investigate the ToxCast™ Phase I data regarding their ability to predict long-term animal toxicity. We published an online platform (*http://iprior.eadmet.com*) that automates large-scale model building and analysis. Moreover, we investigated thousands of models constructed in an effort to predict 61 toxicity endpoints using multiple descriptor packages and hundreds of in vitro assays. We investigated the effect of the pathways and in vitro assays on the resulting model performance. We identified 8 toxicity end points where biologically derived descriptors from in vitro assays or pathway perturbations improved the model prediction ability.
In vivo toxicity end points are generally challenging to model. Only acetylcholinesterase inhibition was possible to readily model with a median balanced accuracy (BA) of all models above 0.6 (BA= 0.76).
We also constructed in silico models to predict the outcome of 144 in vitro assays. This showed better statistical metrics with 40 out of 144 assays having median balanced accuracy of at least 0.6. This suggests that the in vitro datasets have a better modelability than in vivo animal toxicities for the given datasets.

KEYWORDS: computational toxicology, alternative testing, QSAR, ToxCast, REACH, iPRIOR

1 Introduction

The European legislation on chemicals REACH[1] (Registration, Evaluation, Authorization and restriction of chemical) came into effect in the year 2007. The legislation assesses the risk of chemicals and aims to establish safe practices decreasing the impact of chemicals on human health, animal welfare, and the environment.

The legislation aims to collect all available information on a chemical substance in order to to assist the identification of potential sources of hazard and to further convey recommendations on risk management measures through supply chains. The responsibility for the management of substances' risk is transmitted from the regulators to the manufacturers, importers, as well as the traders and users. This raises a huge need to provide accurate information on risk assessment to manufacturers and regulators alike.

In the course of the investigation of information gaps, the European Chemical Agency (ECHA) was established. ECHA aims to manage the databases, which are required to facilitate the information system. Additionally, it coordinates the evaluation of suspicious chemicals and builds and manages a public database for the collected hazard information for consumers and professionals[2]. The benefits from REACH would phase-in gradually as more substances get registered. REACH supports the use of animal testing only as a last resort, but encourages, instead, the justified use of well-established QSAR models, built with respect to the OECD principles, as a valid alternative.

With the evolution in the 'omic' approaches, the *in vitro* profiling of chemicals has been in focus over the previous years, as it appears to offer a potential alternative to long-term *in vivo* animal testing.

Similar studies are conducted worldwide. The U.S. Environmental Protection Agency (EPA) has established a number of programs to profile *in vitro* bioactivity, including ToxCast,[3,4] Tox21,[5,6] and EDSP21 (e1K). EPA shares the aim of significantly reducing animal testing, making testing fast and less costly and of providing characterizations of chemicals, chemical mixtures, and toxicity end points[7]. Figure 1 shows the size of the chemical libraries and number of assays in these projects. Within all mentioned studies ToxCast contains the highest number of *in vitro* assays. It covers about 600 assays. In its Phase I, the program covered 309 chemicals (mostly food pesticides for which thorough animal toxicity studies are available). Figure 1 shows the estimated number of compounds and *in vitro* assays covered in each program. These chemicals come from various sources as shown in Figure 2. This justifies the wide diversity in structural groups, complexity and physicochemical properties.

One of the program's goals is to identify targets or pathways linked to various toxicity endpoints, developing assays for such targets or pathways, and developing predictive toxicological models by using *in vitro* screening data.

The so called "toxicity signatures" are intended to prioritize chemicals for targeted testing and predict possible adverse outcome pathways for chemicals.

Multiple previous studies evaluated the ability of *in vitro* assays for predicting selected *in vivo* endpoints [8–10] and analyzed the biochemical pathways that could be involved with the observed toxicity. Most studies focused on a single *in vivo* toxicity endpoint. A comprehensive analysis of the *in vitro*-to-*in vivo* predictive capability of the ToxCast high-throughput screening effort has been independently presented.[11]

In this study we investigate data from ToxCast Phase I. In particular, we assess the predictive ability of the HTS *in vitro* assays in constructing a toxicity signature both alone and in combination with different *in silico* descriptor packages. We aimed to provide an exhaustive overview of the provided data and its effectiveness and limitations within QSAR studies. We analyze to what extent *in silico* descriptors are able to represent the information in the *in vitro* assays by building *in silico* QSAR models for the prediction of *in vitro* assays output. Certain *in vitro* assays, such as the activation of several nuclear receptors, have already been associated with potential toxicities. The ability to predict such outcome using computational modeling only can save both time and costs.

We deploy a tool for the exploration of huge *in vitro* assay datasets. We perform a large-scale analysis of the ToxCast *in vitro* assays. We analyze their usefulness as biological descriptors for modeling *in vivo* animal toxicity and the possibility to generate *in silico* models to replace the *in vitro* assays.

2 Materials and Methods

2.1 Datasets and data handling

2.1.1 In vitro assays

The Toxminer v17 data were implemented in a local MySQL database and integrated into iPRIOR[12], using Knime.

The database includes information on biochemical pathways, processes, assay-gene, and gene-pathway mappings. Correlations between genes and pathways were collected from Gene Ontology (GO)[13], Kyoto Encyclopedia of Genes and Genomes (KEGG)[14], Ingenuity Pathways analysis (IPA, Ingenuity systems Inc, Redwood city, CA)[15], pathway commons[16], and the OMIM[17] phenotype databases.

The extracted data includes the chemical structure files (in SDF format) of all ToxCast phase I chemicals (309 compounds) in the database. The *in vitro* information consists of 467 assays, some of which evaluate multiple time points, resulting in 669 assay endpoints. It is worth mentioning that the response of the compounds varies significantly across different *in vitro* assay categories. Figure 3 shows the response of the ToxCast phase I chemicals to the 669 endpoints measured. The assays cover nine technologies: cell-free HTS assays; multiplexed transcription reporter; biologically multiplexed activity profiling; high-content cell imaging; multiplexed gene expression; cell-based HTS; phase I and II XME cytotoxicity; real-time cell electronic sensing; and HTS genotoxicity. The assays measure both direct interactions between chemicals and identified receptors and enzymes, as well as downstream events on receptor gene activity or cellular consequence.

EPA database reported a half maximum activity concentration (AC_{50}) or lowest effective concentration (LEC) for assay responses. However, due to the small size of the screening library and comparably low

accuracy under which HTS experiments are conducted, we focused on calculation of classification models. If such models deliver reasonable results, more detailed regression models could be interesting for a further exploration of the underlying endpoints. Therefore, all assay results were discretized into (response/no response) values. The absence of response was reported as a value of 10^6 in the original database. This value was considered as "no response" while any other reported value was considered positive "response".

At a rough estimate, only 7% of the assay/chemical interaction matrix showed a response. Another approach we considered, in terms of data consolidation, is to analyze the liability of a chemical to cause a perturbation in a given pathway, regardless of which gene it affects to cause such perturbation.

Most of the *in vitro* assays show activity for only few compounds or even none at all. Therefore they cannot be modeled with the available data. From the available 669 *in vitro* assay endpoints, only 144 contained 35 or more active hits with the tested compounds. For these endpoints, all 299 concerned compounds were used to build QSAR models. The list of all *in vitro* assays is available from EPA[18].

To calculate chemical-pathway perturbations, 1456 pathways were correlated to 299 chemical structures. We considered the correlation of pathways to their respective genes then investigated whether a compound had a positive hit to any assay associated with these genes. If a chemical shows activity in any assay associated with these genes then it was considered perturbing the investigated pathway. Subsequently we built a chemical/pathway-perturbation matrix, which showed that 14% of potential interactions were positive.

As a single assay correlated with several pathways and vice-versa, the pathway perturbations represent a different re-grouping of the assay data. Such regrouping can result into less sparse datasets and, therefore, is more convenient for application of machine learning methods. One of the goals of this article is to investigate whether such regrouping can improve the predictive models.

2.1.2 Other datasets

Chemical structures and assays resulting from ToxCast Phase II, Tox21, and E1K were uploaded to iPrior and are available for modeling. iPrior uses the concept of tags to identify chemicals. Four tags were created to identify chemicals associated with ToxCast Phase I, Phase II, Tox21 and the E1K projects. Users who are only interested in certain project or class of compounds can set their "area of interest" to certain tags after login. This is intended to limit the scope of their analysis exclusively to associated compounds.

Five substances have no particular structures associated. These are: Cremophor EL, Cornmint oil, Clove leaf oil, Peppermint oil and Anise oil. They were considered unsuitable for QSAR modeling and the associated *in vitro* assay results were excluded from the data upload.

Phase II data included 370462 data points measured on 253 assay endpoints for 1853 chemicals. The modeling of these datasets is outside the scope of this study.

2.1.3 *In vivo* animal studies

The ToxMiner v17 included subset of the toxicity reference database (ToxRefDB) that is relevant for the chemicals of the ToxCast study. The database had results for 461 animal studies conducted. For each toxicity endpoint, *in vivo* toxicity data for chemicals were discretized to a binary outcome (toxic / non-toxic). The same threshold reported with the original database (10^6) was used as the cutoff between positive and negative responses.

ToxRefDB database is a component of the larger ACToR system. It contains summary results of primary toxicology studies submitted to the EPA on pesticide active ingredients[19]. Typically these data have been extracted from EPA Office of Pesticide Programs (OPP) evaluations of studies, based on EPA Office of Prevention, Pesticides and Toxic Substances (OPPTS) harmonized test guidelines. Full details of the collected data has been described in literature[20].
The Toxicity Reference Database (ToxRefDB) has been the primary tool for storing and accessing high-quality toxicology studies and is available online for search and download[21]. ToxRefDB has

characterized thousands of studies using a standardized vocabulary, a uniform structure across study types, and a high level of internal and external quality control (QC) for the extraction of endpoints useful in developing predictive models[20]. Full list of *in vivo* toxicology assays are available in the supplementary materials.

For ToxCast Phase I, animal toxicity studies were linked from ToxrefDB. However, other datasets (i.e ToxCast Phase II, EDSP, Tox21) do not have similar toxicity data readily available. Users are encouraged to contribute such data to the online platform.

2.2 Methods

2.2.1 Interacting with iPrior

iPrior offers a user-friendly web interface (http://iprior.eadmet.com) for interactive data analysis (screenshot shown in Figure 4). It was developed based on OCHEM[22] and QSPR Thesaurus[23] platforms. However, handling large datasets and thousands of QSAR models is more convenient using workflow systems such as Knime[24]. For that, iPrior exposes a number of methods through SOAP web services. These methods allow the user to login, upload data, create properties, create and delete QSAR models, download model statistics, as well as running predictions on the QSAR models.

iPrior implements an xml format that allows users to configure the QSAR modeling tasks with regard to all modeling steps including descriptors calculation, descriptors pre-filtering and configuring the machine learning methods.

Throughout this work, we used different Knime workflows to upload the data, initialize the QSAR modeling on iPrior (as shown in Figure 5) and download the models results.

2.2.2 *In silico* Descriptors

The preprocessing of chemical compounds was conducted using Chemaxon Standardizer, integrated within iPrior workflow. The standardization process included a salt counter-ion removal, charge neutralization, as well as the standardization of nitro groups and aromatic ring representations. For the 3D descriptors, structures were optimized using CORINA[25] which is also integrated within iPrior workflow. The iPrior implementation additionally was used to calculate descriptor values for the *in silico* packages listed in Table 1.

iPrior web platform[12] was used to calculate *in silico* descriptors from different commercial and academic providers (see Table 1). The descriptor values calculated for all structures are available on iPrior.

The calculation for ten structures failed, as these compounds were inorganics, organometalics, mixtures or large macrocyclic compounds. These structures were excluded (see supplementary materials for list of excluded chemicals).

2.2.3 Prefiltering criteria:

All descriptors were pre-filtered before model development. As descriptors with low variance are likely to degrade the performance of certain learning algorithms (in particular those which are distance based), the following pre-filtering criteria were used: first, descriptors that are constant among all compounds, offering no information gain, were removed. Then, normalized descriptors that have variance smaller than < 0.01 were removed. Finally, descriptors with pair-wise Pearson's correlation coefficient (R) > 0.95 were grouped.

The same pre-filtering steps were also applied to biologically derived descriptors (assay results and pathways perturbations) for modeling *in vivo* toxicity endpoints. The list of all *in vitro* assays with the number of their hit compounds is listed in the supplementary materials.

The prefiltering step was also applied within the 5-fold cross-validation loop for any of the machine-learning algorithms used. Thus, the exact numbers of descriptors used could be different for each model.

2.2.4 Machine learning methods

8 machine-learning methods were applied:

k-Nearest Neighbors (kNN) predicts the class membership for a compound using the consensus voting of its nearest k compounds from the training set based on a selected distance metric. The distance metric (e.g. Euclidean or Manhattan) is usually defined in the descriptor space. We used the Euclidean distance calculated using normalized descriptors (mean 0 and standard deviation 1). Naturally, kNN works well only in balanced training sets[26,27].
As the parameter k influences the decision of class membership, the number of nearest neighbors that provided the highest accuracy of classification was calculated following a systematic search in the range (1, 100).

Associative neural networks (ASNN)[28,29]: belongs to the family of multilayered perceptron[30] neural networks. It can be represented as a multilayered graph in which all nodes of certain layer are connected to the nodes of the previous one. The resulting class membership is the output of the single neuron in the last layer of the network. ASNN uses kNN over the space of ensemble predictions. This allows for a local correction for the ensemble of neural networks. The distance is based on the correlation between the vectors of predicted samples by the networks of the ensemble. The configuration of the algorithm was kept to OCHEM defaults (i.e., 3 neurons in the hidden layer, 1000 iterations, using model ensemble size of 64, the method for neural network training was SuperSAB)

C4.5 decision tree (J48): is a decision tree classifier based on the concept of entropy gain[31]. The tree nodes are optimized to split the molecule sets most effectively between the binary classes. The criterion for this optimization is choosing the descriptor that results into maximum normalized information gain (entropy difference). J48 is a Java implementation of C4.5 decision tree in the statistical software WEKA[32,33]. The default parameters provided by WEKA were used with no further optimization.

Multiple Linear Regression Analysis (MLRA): Regression methods detect continuous correlation between the descriptor space and the AHR class membership. It predicts the activity as a function of an optimal linear combination of descriptors, which is selected to minimize the training set error. Studies reported that MLRA is prone to over-fitting[34,35] (misinterpretation due to the use of large number of intercorrelated descriptors). Thus prefiltering of correlated descriptors is necessary.

The MLRA method selected uses stepwise variable selection. It eliminates a single descriptor on each step. The descriptor selected for elimination when its regression coefficient is insignificantly different from zero according to the t-test. The method has only one parameter, ALPHA, which corresponds to the p-value of variables to be kept for the regression. We used an ALPHA value of 0.05.

Fast Stagwise Multiple Linear Regression (FSMLR)[36]: is a procedure for stagewise building of linear regression models by means of greedy descriptor selection.

Partial Least Squares (PLS): Performs orthogonal transformation of the descriptor space[37]. The latent variables are ranked by their variance in the descriptor space as well as their correlation to the class membership. The number of latent variables was optimized automatically using 5-fold cross-validation on the training set.

Random Forests (RF)[38]: is a meta-learning method built on the concept of random decision tree algorithm[39] using bootstrap sampling among all training set molecules. Multiple trees are created and the final class membership results from the consensus voting of individual trees.
The method used was a Java implementation of random forests in the statistical software WEKA[32,33]. Each RF model was built using 10 trees. The default parameters provided by WEKA were used with no further optimization.

Support Vector Machine (SVM)[40]: uses the LibSVM program. The SVM method has two important configurable options: the SVM type (ε -SVR and μ -SVR) and the kernel type (linear, polynomial, radial basis function, and sigmoid). Classic ε -SVR and radial basis function kernels were used. The other SVM parameters, namely cost C and width of the RBF kernel, were optimized using default grid search, which was performed according to the LibSVM manual.

2.2.5 Performance measures used to compare models

Several measures were analyzed to estimate the accuracy of models. These measures include: total accuracy (ACC), balanced accuracy (BAC), Matthews correlation coefficient (MCC), positive predictive value (PPV), sensitivity, and specificity. Table 2 shows the equations for these measures. Due to the unbalanced nature of the datasets, balanced accuracy was used throughout the text as the primary measure for comparing models.

2.2.6 Modeling *in vivo* animal toxicity

Different feature combinations were used for modeling. These are the 11 *in silico* descriptor packages as well as 6 biological-derived features. These are: the ToxCast *in vitro* assay, the pathway perturbations as described before, the combination of both and finally combining CDK descriptors (as an example of a widely used *in silico* descriptors package) with each of the three features.

Out of the 461 available animal studies, only 61 showed toxicity for 35 or more chemicals, which was used as a tentative threshold for conducting proper 5-fold cross validation. For every endpoint the total number of tested compounds was between 234-251. Animal toxicity studies were conducted in rats, rabbits, and mice. For each study only one animal species was used.

2.2.7 Modeling *in vitro* assays

An interesting exploration was to figure out to which extent could *in silico* descriptors represent the information represented in the *in vitro* assays. To investigate this, we evaluated approaches to model the *in vitro* assays using the 11 *in silico* packages listed in Table 1.

Although assay results were available for all ToxCast Phase I compounds. Only 144 assay endpoints showed positive response for 35 or more chemicals, which was used as a tentative threshold for conducting proper 5-fold cross validation. These endpoints were therefore used for modeled.

3 Results and Discussion

3.1 Accessing results on iPrior

All models are associated with a unique identification number (model id). That id can be used to access the model profile page (see Figure 6). To access a certain model, users visit: http://iprior.eadmet.com/model/[modelID] replacing [modelID] with the model identification number.

The profile page lists, besides the model name and property predicted, the algorithm and descriptors used, pre-filtering parameters as well as the model statistics. From this page, users have also access to the applicability domain graphs as shown in Figure 7. These graphs are automatically calculated (whenever applicable) based on the distance-to-model (DM) approach [41]. Currently, several DMs are supported: the standard deviation of an ensemble of models (STDEV), the correlation in the space of models (CORREL)[42] and Mahalanobis distance (LEVERAGE).

Model quality can be judged through the statistical parameters presented in the model profile page. Figure 8 shows an example of a poor-quality model. Users can notice both accuracy and balanced accuracy are below 0.5.

Whenever Users are satisfied with the quality of the model they can apply it to new compounds. They get the option to draw a chemical structure directly through the web browser, select from a previous dataset or upload their own structure file in SDF format.

Modeling results and statistics can also be queried using Workflow system such as Knime as shown in Figure 9.

3.2 Modeling *in vivo* animal toxicity

In total 8 machine-learning approaches were applied to 17 feature combinations (see Table 1) to model the 61 *in vivo* toxicity endpoints resulting in 8296 QSAR models. The selected *in vivo* animal toxicities for modeling together with their respective number of toxic compounds are listed in the supplementary materials. All models were built based on a 5-fold cross- validation.

Figure 10shows the balanced accuracy of all 8296 models built. Different statistical parameters for all the models including: sensitivity, specificity, balanced accuracy, as well as Matthews's correlation coefficient (MCC) are included in the supplementary materials.

The balanced accuracy of the five best predicted animal toxicity studies are provided in Table 3. The ranking was based on the maximum achievable balanced accuracy of all models (144) generated for the respective end point. The machine learning algorithms that contributed to the best models were differed widely between the cases. In some cases it was better to have a linear algorithm while in others the non-linear algorithms predominated. This could be related to the complexity of the end point and the descriptors involved.

3.2.1 Understanding significant features using ToxAlerts

As shown in Table 3, rat acetylcholinesterase inhibition was one of the most predictable endpoints, with the balanced prediction accuracy reaching up-to 90%. The analysis below concerns this endpoint. It was used as an example to understand the reason behind the success in modeling some end points. To analyze the most significant features contributing to toxicity, we used the "Set Compare" and "ToxAlerts" tools previously developed within OCHEM platform[43–45]. ToxAlerts is an open Web-based platform for uploading and storing structural alerts published in scientific literature with a capability to virtually screen compound libraries against these alerts to flag toxic chemicals or compounds with potential adverse drug reactions. The alerts are uniquely identified SMARTS patterns collected from literature sources. The database contains more than 2000 alerts from more than 25 publications for carcinogenicity, mutagenicity, skin sensitization, acute aquatic toxicity, and potential idiosyncratic drug toxicity. It has previously been used to identify small-molecule frequent hitters from AlphaScreen high throughput screens as well[46].
When the 2 sub-sets of the rat acetylcholinesterase inhibition compounds (toxic vs. non-toxic) were compared using ToxAlerts[43], many significant toxic groups were detected. The three most significant alerts are shown in Table 4. Indeed, acetylcholinesterase inhibition is the main mechanism of action by which Organophosphorus insecticides perform their function. Only one non-toxic compound was a phosphorus derivative. This simple scaffold is easy to capture for any descriptor package that accounts for fragments or atom counts while becomes harder for *in vitro* assays to indirectly capture the presence of that scaffold. Table 4 also shows the enrichment factor and the p-value for the significance of each SMARTS pattern detected.

3.2.2 Detecting significant *in vitro* assays using SetCompare

To detect which *in vitro* assays or pathways perturbations can better act as a biological descriptor for building a toxicity signature, the "Set Compare" tool on iPrior was directly applied. The Acetylcholinesterase inhibition was used as an example end point for the analysis. The toxic vs. non-toxic sets of compounds were compared. Table 5 shows the most significant *in vitro* assays. Indeed, the impact on both, rat and human, Acetylcholinesterase (AChE) were the most significant *in vitro* assays.

For each toxicity end point, we selected the individual model that showed the highest balanced accuracy. We then ranked different algorithms and descriptor packages according to how many times they would contribute to such models.

Table 7 shows the ranking of different descriptor packages in their success to achieve the best predictive model regarding the balanced accuracy. It also shows the number of toxicity end points for which the descriptor package contributed to its best model.
Generally, ALOGPS + OESTATE descriptors were ranked best while *biological descriptors* performed worst.

It is worth noticing that biological descriptors outperformed *in silico* descriptors for the prediction of 5 toxicity end points. Also, Biological descriptors in combination with *in silico* descriptors contributed to achieving the highest balanced accuracy for 3 other end points. Table 6 lists these end points. It also presents the balanced accuracies of the biological descriptors, *in silico* descriptors alone and the combination of them. In three cases, it was the combination of both that resulted in the best model.

This might be because each kind of descriptors might encode for different information related to the chemical structure or the *in vivo* target.
In 5 cases, the pathways, either alone, in combination with the assays or in combination with the *in silico* descriptors significantly improved the model prediction ability (Table 6). This suggests that the re-arrangement of *in vitro* assay outcomes in the form of pathways perturbation provided extra information for modeling these toxicity endpoints.
Most of the end points were either related to a productive end point or a pregnancy developmental one.

Table 8 compares different algorithms on their performance. In general, comparably simple method, such as the FSMLR, MLRA and kNN approach shows better performance compared to non-linear high-resolution methods, such as random forests and neural networks. This might indicate the absence of significant non-linear dependencies.

3.3 Modeling *in vitro* assays

The same learning algorithms were used as for the *in vivo* animal experiment modeling. The *in silico* descriptors from the 11 descriptor packages were applied. In total 12672 models were built on 144 endpoints applying 8 machine learning approaches to nine feature combinations). The selected *in vitro* end points for modeling are listed in the supplementary materials. All models were built based on a 5-fold cross- validation.

Regarding the best performing descriptors, the observations were different from the case with the *in vivo* experiments. In this case, there was no significant difference between the performances of different descriptor packages as shown in

Table 7. Analogously, Table 8 provides a comparison of different machine learning algorithms. Also in this case all algorithms performed comparably well.

Figure 11 shows the balanced accuracy of all 12672 models built to predict the *in vitro* assays. As with the animal toxicity modeling, different statistical parameters are reported in the supplementary materials. In comparison to the prediction of *in vivo* experiments a significant improvement in accuracy is observable. For numerous assays, such as those listed in Table 9, we were able to build predictive models exceeding a balanced accuracy of 0.8. Table 9 lists the statistics for the five best-predicted *in vitro* assays based on the best-achievable balanced accuracy of their respective 88 models built per end-point. Among the 144 in –vitro assays, 68 endpoints showed balanced accuracy of >0.6. Among the most successful assays to be modeled are those related to change in expression of different isoforms of the liver metabolizing enzymes CYP450. This also agrees with previous *in silico* studies that reported success in building *in silico* QSAR models for prediction of CYP450 expression change of different isoforms[47–49].

Modelability of the datasets:
The predictivity of QSAR models is directly influenced by various data set characteristics [50,51](e.g., size, chemical diversity, activity distribution, presence of activity cliffs, etc.) as well as the modeling process itself. Previous study investigated ToxCast Phase I data with regard to its modelability showing that the presence of too many activity cliffs compare to the dataset size makes it not suitable for QSAR modeling [52]. We used the median balanced accuracy of all models built for each individual end point to compare its relative ease of modeling.

For *in vivo* endpoints, with one exception, none of the end points had a median balanced accuracy of above 0.6 (with 68% confidence interval).

The chronic rat acetylcholinesterase inhibition stands as a clear exception for an end point that is easy to model. The median balanced accuracy for all models exceeds 0.75.

For *in vitro* end points, the provided statistics reveal that 4 end points have median balanced accuracy above 0.68 (with 68% confidence interval). Comparatively, *in vitro* assays where easier to model than *in vivo* toxicity end points.

3.4 Running predictions

iPrior is a public online tool. It allows users to run predictions on any of the QSAR models created in this study. The supplementary materials contain the model ids for all generated models. Users are encouraged to investigate the model profiles, applicability domains and run predictions using their own chemical structures. Models can be accessed through their public identification number by visiting:
http://iprior.eadmet.com/model/[ID-number]
Replacing [ID -number] with the provided public ID.

4 Conclusion

Analysis of Phase I compounds shows that, the dataset is challenging to build predictive toxicity models for replacement of animal testing with the exception of Acetylcholinestrase inhibition. However, comprehensive modeling with multiple machine learning algorithms and descriptors shows relative success for selected end points (Table 3). Low modelability of most end points could be due to the limited chemical diversity of the dataset, consisting mainly of insecticides and pesticides.

Biological descriptors derived from the *in vitro* profiling of chemicals had a significant improvement in the models' predictive ability in some cases (Table 6). The use of *in vitro* assays, as biological descriptors, showed better modeling performance for a handful of toxicity endpoints than use of *in silico* descriptors. Also the regrouping of the *in vitro* assay responses in the form of pathway perturbations significantly improved the predictivity of some toxicity end points.

We identified several *in vivo* endpoints, which appear to be predictive with balanced accuracy higher than 0.75 (examples listed in Table 3). Furthermore, multiple *in vitro* assays showed a high balanced accuracy (>0.8) (Table 9) when modeled by *in silico* descriptors. This might be due to the fact that each *in vitro* assay is typically measuring a small number of interacting genes and pathways, which is insufficient, when considering the more complex requirements needed to model a toxicity phenotype. This enables an improved approach towards *in silico* modeling of toxicity in general.

Many challenges remain in place: first of all, the use of a statistical approach, such as QSAR modeling, requires a considerable amount of data. The comparably low number of training instances limits the possibility to model the data in an appropriate way. Secondly, the *in vitro* representation could be too simple to address the complexity of the interactions in the organism. Properties, such as bioavailability and biotransformation play a significant role in terms of the toxic effect of a compound. Thirdly and finally, it is possible that the assays conducted are not enough to capture biochemical events on the molecular level that can describe the pathways responsible for toxicity.

A consequence arising from this should be the careful investigation and analysis of potentially useful *in vitro* assays in terms of specific toxicity endpoints, as well as the identification of those *in vitro* assays, which enable proper QSAR modeling.

With that taken into consideration, ToxCast Phase I still provided useful overview of the chemical initiating events that could be useful for further investigation with a higher number compounds. For example, many assays may be redeemed unnecessary in future tests, as they were focused on promiscuous dormant endpoints.

As more information is being gathered from Chemical providers through programs like REACH and ECHA Tox21 projects, QSAR modeling will play more significant role.

We deployed the public platform iPrior[12] (Figure 4) . The platform is freely accessible for the non-commercial use of the academic community.
The required workflows and modeling infrastructure is now in place that could assist scientists in developing predictive biological signatures. Such infrastructure is available for the analysis of future data releases.

As more data become available with the progress of the next phases of ToxCast and similar projects, it would be possible to build statistical models that covers a wider applicability domain and better support the prediction of toxicity and therefore reduce the number of animal experiments. For that, a

central repository of public data that is open to all researchers was created. iPrior (http://iprior.eadmet.com) currently hosts data from ToxCast, Tox21, e1K projects. It is open to researchers to upload more data or contribute their descriptor packages. iPrior supports the full cycle of QSAR research online. It can calculate descriptors, prefilter parameters, run machine-learning algorithms, cross-validation and bootstrap aggregation. Final models are stored and can be shared among users. Predictions can then be run for new molecules.

Finally, *in vitro* assays for chemical profiling remains a useful investigation and exploratory tool. We used the "Set Compare" utility to detect the most significant *in vitro* assays for the identification of toxicity endpoints (Table 5). It provides a tool for suggesting the mechanism of toxicity of chemicals. Previous study[11] showed similar results and suggested that *in vitro* profiling could be useful for the prioritization of compounds, rather than the replacement of animal testing.

5 Acknowledgement

This study was partially supported by the FP7 MC ITN project Environmental Chemoinformatics (ECO), grant agreement No. 238701 and the BMBF GO-Bio grant number 0315647.

6 Supportive/Supplementary Material

Supplementary 1: List of *in vivo* end points from ToxCast / ToxrefDB, their respective total number of hits and whether it was selected for modeling.
Supplementary 2: List of ToxCast Phase I chemicals excluded from modeling due to failed descriptors calculation.
Supplementary 3: List of *in vitro* assay end points, their respective total number of hits and whether it was selected for modeling.
Supplementary 4: Detailed statistical parameters for 8296 models constructed using 136 algorithm/features combinations for each of the 61 in vivo toxicological end point from the Toxicity reference database
Supplementary 5: Detailed statistical parameters for 12672 models constructed using 88 algorithm/in silico descriptors combinations for each of the 144 in vitro assay endpoints from the ToxCast database.

7 References

[1] REACH - Registration, Evaluation, Authorisation and Restriction of Chemicals http://ec.europa.eu/enterprise/sectors/chemicals/reach/index_en.htm (accessed Sep 21, 2013).

[2] REACH - Chemicals - Environment - European Commission http://ec.europa.eu/environment/chemicals/reach/reach_en.htm (accessed Sep 8, 2014).

[3] Judson, R.S.; Houck, K.A.; Kavlock, R.J.; Knudsen, T.B.; Martin, M.T.; Mortensen, H.M.; Reif, D.M.; Rotroff, D.M.; Shah, I.; Richard, A.M.; Dix, D.J. In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. *Environ. Health Perspect.*, **2010**, *118*, 485–492.

[4] 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.

[5] Raymond Tice Robert Kavlock, Ph.D., and Christopher Austin, M.D., P.D.; Tice, R.; Kavlock, R.; Austin, C. The U.S. "Tox21 Community" and the Future of Toxicology.

[6] Betts, K.S. Tox21 to Date: Steps toward Modernizing Human Hazard Characterization. *Environ. Health Perspect.*, **2013**, *121*, A228.

[7] US EPA, O. of P.P. Overview of National Research Council Toxicity Testing Strategy | Pesticides | US EPA.

[8] Martin, M.T.; Knudsen, T.B.; Reif, D.M.; Houck, K.A.; Judson, R.S.; Kavlock, R.J.; Dix, D.J. Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening. *Biol. Reprod.*, **2011**, *85*, 327–339.

[9] Shah, I.; Houck, K.; Judson, R.S.; Kavlock, R.J.; Martin, M.T.; Reif, D.M.; Wambaugh, J.; Dix, D.J. Using Nuclear Receptor Activity to Stratify Hepatocarcinogens. *PLoS One*, **2011**, *6*, e14584.

[10] Kleinstreuer, N.C.; Judson, R.S.; Reif, D.M.; Sipes, N.S.; Singh, A. V; Chandler, K.J.; Dewoskin, R.; Dix, D.J.; Kavlock, R.J.; Knudsen, T.B. Environmental Impact on Vascular Development Predicted by High Throughput Screening. *Environ. Health Perspect.*, **2011**.

[11] Thomas, R.S.; Black, M.B.; Li, L.; Healy, E.; Chu, T.-M.; Bao, W.; Andersen, M.E.; Wolfinger, R.D. A Comprehensive Statistical Analysis of Predicting in Vivo Hazard Using High-Throughput in Vitro Screening. *Toxicol. Sci.*, **2012**, *128*, 398–417.

[12] iPrior - Prioritization and estimation of toxicity of chemcial compounds http://iprior.eadmet.com/home/show.do (accessed Sep 24, 2013).

[13] Gene Ontology Documentation http://www.geneontology.org/GO.contents.doc.shtml (accessed Sep 21, 2013).

[14] KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg/ (accessed Sep 21, 2013).

[15] Ingenuity IPA - Integrate and understand complex 'omics data http://www.ingenuity.com/products/ipa (accessed Sep 21, 2013).

[16] Pathway Commons http://www.pathwaycommons.org/about/ (accessed Sep 21, 2013).

[17] Boyadjiev, S.A.; Jabs, E.W. Online Mendelian Inheritance in Man (OMIM) as a Knowledgebase for Human Developmental Disorders. *Clin. Genet.*, **2000**, *57*, 253–266.

[18] US EPA, O. of W.C. and O. of E.I. EPA ACTOR Downloads http://actor.epa.gov/actor/faces/Download.jsp (accessed Sep 21, 2013).

[19] Martin, M.T.; Houck, K.A.; McLaurin, K.; Richard, A.M.; Dix, D.J. Linking Regulatory Toxicological Information on Environmental Chemicals with High-Throughput Screening (HTS) and Genomic Data. *Toxicol. CD-An Off. J. Soc. Toxicol.*, **2007**, *96*, 219–220.

[20] Martin, M.T.; Judson, R.S.; Reif, D.M.; Kavlock, R.J.; Dix, D.J. Profiling Chemicals Based on Chronic Toxicity Results from the US EPA ToxRef Database. *Environ. Health Perspect.*, **2009**, *117*, 392.

[21] US EPA, O. of W.C. and O. of E.I. ToxRefDB (Toxicity Reference Database).

[22] Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A.K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V; Tanchuk, V.Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin II; Palyulin, V.A.; Radchenko, E. V; Welsh, W.J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided. Mol. Des.*, **2011**, *25*, 533–554.

[23] Brandmaier, S.; Peijnenburg, W.; Durjava, M.K.; Kolar, B.; Gramatica, P.; Papa, E.; Bhhatarai, B.; Kovarich, S.; Cassani, S.; Rahmberg, M.; Öberg, T.; Jeliazkova, N.; Golsteijn, L.; Comber, M.; Charochkina, L. The QSPR-THESAURUS : The Online Platform of the CADASTER Project. **2014**, 1–12.

[24] Guide, K.Q.; Screen, W.; Status, N.; Flow, E.; Nodes, A.; Nodes, C.; Nodes, C.; Nodes, E.; Voyage, Y.O.; Workbench, K.; Guide, U.; Views, A.; Projects, W.; Nodes, F.; Repository, N.; Description, N.; Gui, K.; Key, M.; Editor, W.; Options, N.; All, E.; View, O.; View, O.O.; All, C.; Custom, E.; Name, N.; Knime, C.; Page, G. KNIME Quickstart Guide. 1–27.

[25] Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 1000–1008.

[26] Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When Is "nearest Neighbor" Meaningful? In *Database Theory—ICDT'99*; Springer, **1999**; pp. 217–235.

[27] Aha, D.W.; Kibler, D.; Albert, M.K. Instance-Based Learning Algorithms. *Mach. Learn.*, **1991**, *6*, 37–66.

[28] Tetko, I. V. Associative Neural Network. *Neural Process. Lett.*, **2002**, *16*, 187–199.

[29] Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.*, *42*, 717–728.

[30] Rosenblatt, F. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*; Cornell Aeronautical Laboratory, **1957**.

[31] Quinlan, J.R. *C4. 5: Programs for Machine Learning*; Morgan kaufmann, **1993**; Vol. 1.

[32] Holmes, G.; Donkin, A.; Witten, I.H. WEKA: A Machine Learning Workbench. In; **1994**; pp. 357 –361.

[33] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.*, **2009**, *11*, 10–18.

[34] Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. Multi-Way Principal Components-and PLS-Analysis. *J. Chemom.*, **1987**, *1*, 41–56.

[35] Eriksson, L.; Johansson, E. Multivariate Design and Modeling in QSAR. *Chemom. Intell. Lab. Syst.*, **1996**, *34*, 1–19.

[36] Zhokhova, N.I.; Baskin, I.I.; Palyulin, V.A.; Zefirov, A.N.; Zefirov, N.S. Fragmental Descriptors with Labeled Atoms and Their Application in QSAR/QSPR Studies. In *Doklady Chemistry*; **2007**; Vol. 417, pp. 282–284.

[37] Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.*, **2001**, *58*, 109 – 130.

[38] Breiman, L. Bagging Predictors. *Mach. Learn.*, **1996**, *24*, 123–140.

[39] Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC press, **1984**.

[40] Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, **2011**, *2*, 27.

[41] Tetko, I. V; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.*, **2008**, *48*, 1733–1746.

[42] Tetko, I. V; Bruneau, P.; Mewes, H.-W.; Rohrer, D.C.; Poda, G.I. Can We Estimate the Accuracy of ADME-Tox Predictions? *Drug Discov. Today*, **2006**, *11*, 700–707.

[43] Sushko, I.; Salmina, E.; Potemkin, V. a; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.*, **2012**.

[44] Vorberg, S.; Tetko, I. V. Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). *Mol. Inform.*, **2014**, *33*, 73–85.

[45] Tetko, I. V; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A.E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of Dimethyl Sulfoxide Solubility Models Using 163,000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. *J. Chem. Inf. Model.*, **2013**, *53*, 1990–2000.

[46] Schorpp, K.; Rothenaigner, I.; Salmina, E.; Reinshagen, J.; Low, T.; Brenke, J.K.; Gopalakrishnan, J.; Tetko, I. V; Gul, S.; Hadian, K. Identification of Small-Molecule Frequent Hitters from AlphaScreen High-Throughput Screens. *J. Biomol. Screen.*, **2014**, *19*, 715–726.

[47] Roy, K.; Roy, P.P. QSAR of Cytochrome Inhibitors. **2009**.

[48] Lewis, D.F. V; Modi, S.; Dickins, M. Structure-Activity Relationship for Human Cytochrome P450 Substrates and Inhibitors. *Drug Metab. Rev.*, **2002**, *34*, 69–82.

[49] Novotarskyi, S.; Sushko, I.; Korner, R.; Pandey, A.K.; Tetko, I. V. A Comparison of Different QSAR Approaches to Modeling CYP450 1A2 Inhibition. *J. Chem. Inf. Model.*, **2011**, *51*, 1271–1280.

[50] Fourches, D.; Tropsha, A. Using Graph Indices for the Analysis and Comparison of Chemical Datasets. *Mol. Inform.*, **2013**, *32*, 827–842.

[51] Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.*, **2010**, *29*, 476–488.

[52] Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data Set Modelability by QSAR. *J. Chem. Inf. Model.*, **2014**, *54*, 1–4.

[53] Tetko, I. V; Tanchuk, V.Y.; Villa, A.E.P. Prediction of N-Octanol/water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1407–1421.

[54] Tetko, I. V; Tanchuk, V.Y.; Kasheva, T.N.; Villa, A.E. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1488–1493.

[55] Hall, L.H.; Kier, L.B.; Brown, B.B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 1074–1080.

[56] Calculator Plugins « ChemAxon – cheminformatics platforms and desktop applications https://www.chemaxon.com/marvin/help/calculations/calculator-plugins.html (accessed Sep 28, 2013).

[57] Aires-de-Sousa, J.; Gasteiger, J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 369–375.

[58] Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solovev, V.; Hoonakker, F.; Tetko, I. V; Marcou, G. ISIDA-Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided. Drug Des.*, **2008**, *4*, 191–198.

[59] Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo-and Bioinformatics. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 493–500.

[60] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley. com, **2009**.

[61] Cherkasov, A.; Ban, F.; Santos-Filho, O.; Thorsteinson, N.; Fallahi, M.; Hammond, G.L. An Updated Steroid Benchmark Set and Its Application in the Discovery of Novel Nanomolar Ligands of Sex Hormone-Binding Globulin. *J. Med. Chem.*, **2008**, *51*, 2047–2056.

[62] Potemkin, V.A.; Grishina, M.A. A New Paradigm for Pattern Recognition of Drugs. *J. Comput. Aided. Mol. Des.*, **2008**, *22*, 489–505.

[63] Grishina, M.A.; Bartashevich, E. V; Potemkin, V.A.; Belik, A. V. Genetic Algorithm for Predicting Structures and Properties of Molecular Aggregates in Organic Substances. *J. Struct. Chem.*, **2002**, *43*, 1040–1044.

[64] Potemkin, V.A.; Pogrebnoy, A.A.; Grishina, M.A. Technique for Energy Decomposition in the Study of "Receptor-Ligand" Complexes. *J. Chem. Inf. Model.*, **2009**, *49*, 1389–1406.

[65] Thormann, M.; Vidal, D.; Almstetter, M.; Pons, M. Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names. *Open Appl. Informatics J.*, **2007**, *1*, 28–32.

[66] Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J.P. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *J. Phys. Chem. A*, **2002**, *106*, 7895–7901.

[67] Bultinck, P.; Langenaeker, W.; Carbó-Dorca, R.; Tollenaere, J.P. Fast Calculation of Quantum Chemical Molecular Descriptors from the Electronegativity Equalization Method. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 422–428.

[68] ADRIANA.Code - Calculation of Molecular Descriptors | Inspiring Chemical Discovery http://www.molecular-networks.com/products/adrianacode (accessed Sep 28, 2013).

8 Figures

Legends for figure

Figure 1 Different screening programs managed by the EPA and its partners. The Toxcast program has the most comprehensive number of assays while the Tox21 project includes the most diverse set of chemicals (8300). ToxCast phase III will extend the chemical library of ToxCast by 1000 new compounds and an additional 200 assays.

Figure 2 Inventory sources for ToxCast Phase I & II chemicals. Phase I & Phase II covers 1060 chemical compounds, EDSP21 (e1k) adds another 800 compounds (total: 1860). 2806 total overlaps across 16 diverse inventories. GRAS: Food and Drug Administration (FDA) - Generally Recognized as Safe. MPV: Medium Production Volume,, FDA CFSAN: Center for Food Safety and Applied Nutrition, EDSP: Endocrine Disruptor Screening Program, NTP: National Toxicology Program, TRI: Toxics Release Inventory, HPV: High Production Volume, IRIS: Integrated Risk Information System

Figure 3. Histogram showing count of chemicals showing positive assay and pathway hits for 309 compounds of ToxCast Phase I. The assay data (blue bars) is very sparse - most chemicals affect only a few assays. Grouping by assays by affected pathways (red bars) allows to retrieve a dataset that is less sparse and, therefore, more convenient for machine learning algorithms.

Figure 4 Screenshot of the iPrior homepage showing different in vitro assays for which data are available as well as an excerpt of the published models available for users to run predictions against.

Figure 5 Knime workflow showing the QSAR models building process on iPrior. Different loops iterate over the model configuration XML files and end points to model. Overall 20968 were built.

Figure 6 Model profile page for a good performing model showing (1) model name (2) model id (3)the predicted end point (4) the machine-learning algorithm used (5) The configuration for the learning algorithm and the pre-filtering parameters (6) The model's accuracy, balanced accuracy, Matthew's Correlation Coefficient and AUC of the ROC (7) The ROC curve (8) model confusion matrix showing Hit rate and Precision (9)different tools allowing model statistics download, model replication, exporting model configuration or analyzing the data matched molecular pairs.

Figure 7 The applicability domain graph for the above model showing distance to model in terms of standard deviation of the ASNN ensemble (x-axis) and model accuracy (y-axis)

Figure 8 Model profile page for a bad performing model showing (1) model name (2) model id (3)the predicted end point (4) the machine-learning algorithm used (5) The configuration for the learning algorithm and the pre-filtering parameters (6) The model's accuracy, balanced accuracy, Matthew's Correlation Coefficient and AUC of the ROC (7) The ROC curve (8) model confusion matrix showing Hit rate and Precision (9)different tools allowing model statistics download, model repliaction, exporting model configuration or analyzing the data matched molecular pairs.

Figure 9 Parital Knime workflow showing steps to query iPrior for QSAR model ID's, requesting model status, filtering by ready models and downloading their statistics

Figure 10. Plot showing the variation in the balanced accuracy for the 8296 models constructed using 136 algorithm/features combinations for each of the 61 in vivo toxicological end point from the Toxicity reference database. The upper and lower edges of the line represents the maximum and minimum balanced accuracy achieved; respectively. End points are sorted by the maximum achievable balanced accuracy. The triangle shows the median balanced accuracy among the models irrespective of the in silico or biological descriptors or the machine-learning algorithm used. Full statistics of each model is available in the supplementary materials The endpoints names are shown on the x-axis ordered alphabetically based on the format in ToxRefDB database: study type_species_organ_effect_category. The full list of endpoints and their description is available from EPA website[18] . Study type: DV, developmental; CHR, chronic; MGR, multigenerational. Species: Rt, rat; Rb, rabbit; Ms, mouse. Effect and category: Mat, maternal; GL-Mt, general maternal; Dev,

developmental; PregRel, pregnancy related; PregLoss, pregnancy loss; AnyLes, any lesion; Skl, skeletal; PreneoplastLes, preneoplastic lesion; GenFetal, general fetal; Prolif- eratLes, proliferative lesion; WghtReg, weight reduction; NeoplastLes, neoplastic lesion; Reproduct, reproductive; ThyroidGlnd, thyroid gland; ReproductTract, reproductive tract; Perform, performance; Cholinester, cholinesterase; Inhibit, inhibition.

Figure 11 Plot showing the variation in the balanced accuracy for the 12672 models constructed using 88 algorithm/in silico descriptors combinations for each of the 144 in vitro assay endpoints from the ToxCast database. The upper and lower edges of the line represents the maximum and minimum balanced accuracy achieved; respectively. End points are sorted by the maximum achievable balanced accuracy. The triangle shows the median balanced accuracy among the models irrespective of the in silico descriptor package or the machine learning algorithm used. Full statistics of each model is available in the supplementary materials. ACEA: ACEA - Real-time Cell Electronic Sensing; ATG: Attagene - Transcription factor assays; BSK: BioSeek - Cell-based protein level assays; CLM: Cellumen - Cell imaging assays; CLZD: CellzDirect - Transcription assays; NCGC: NCGC - nuclear receptor assays; NVS: Novascreen / Caliper - receptor binding and enzyme inhibition assays; Solidus: Solidus - P450 vs. cytotoxicity assays

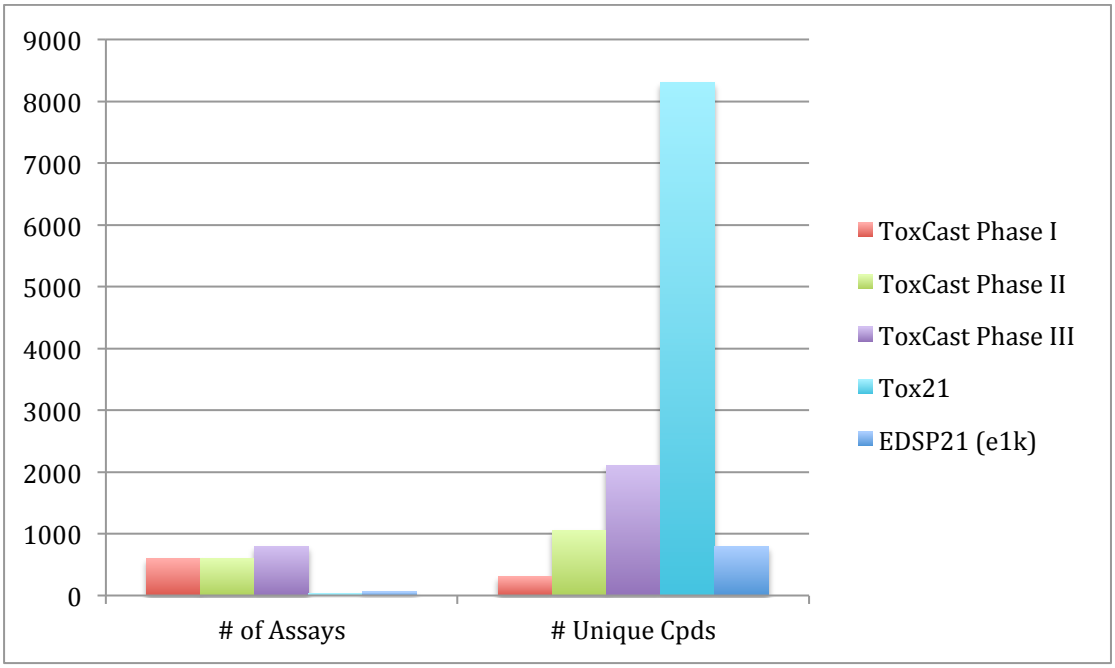


Figure 1

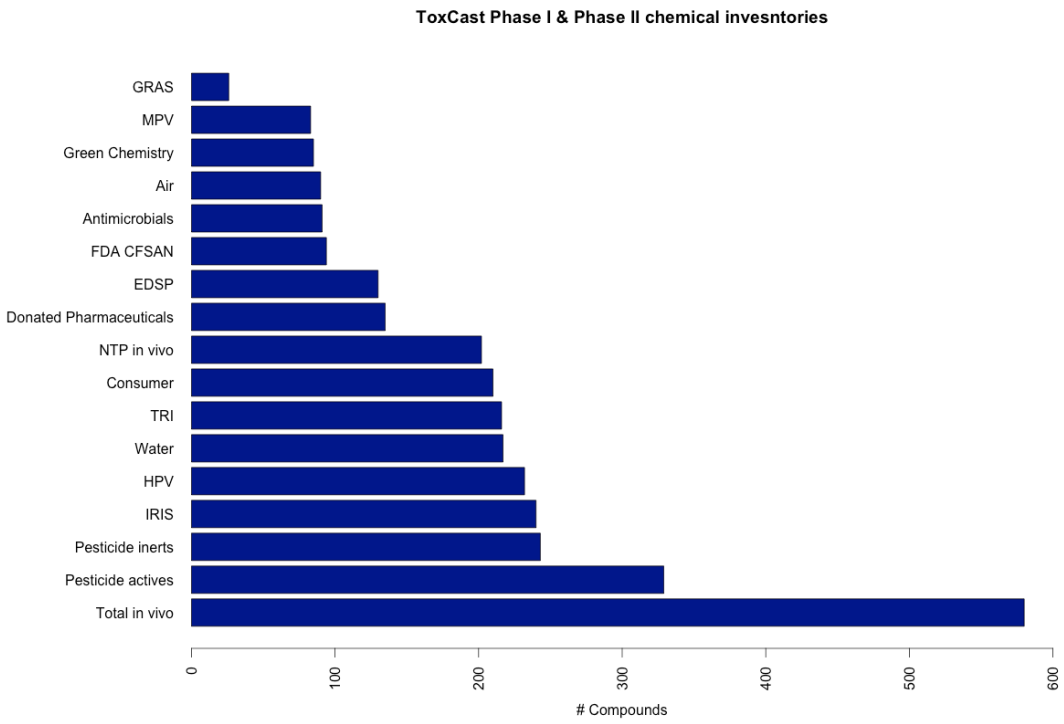


Figure 2

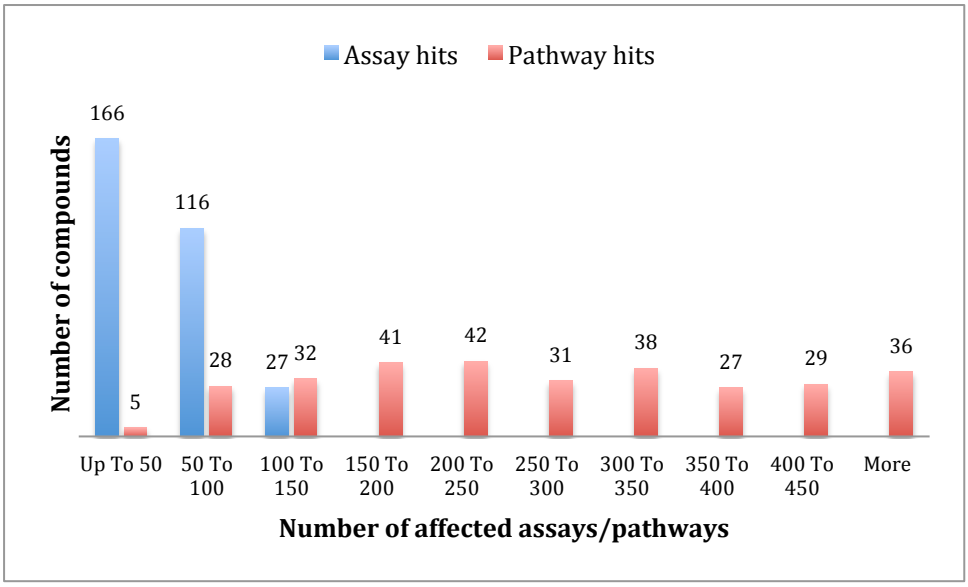


Figure 3.

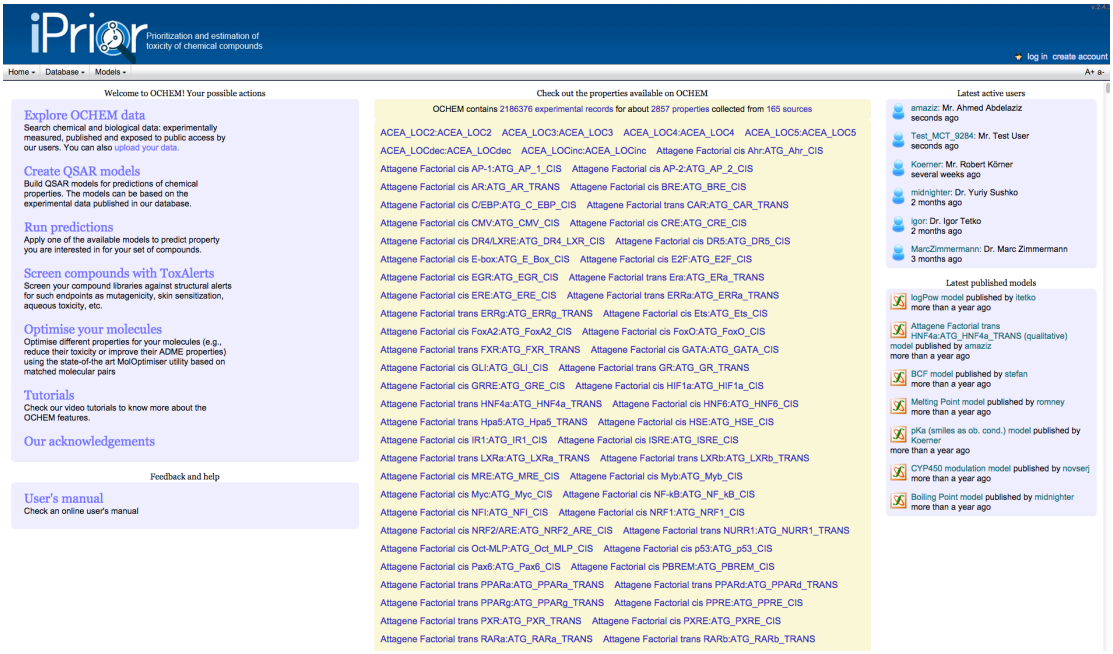


Figure 4

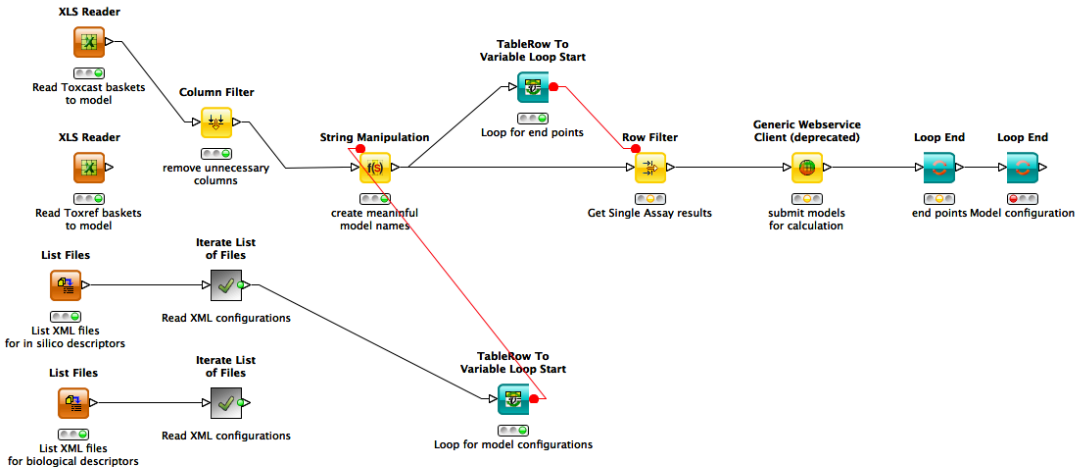


Figure 5

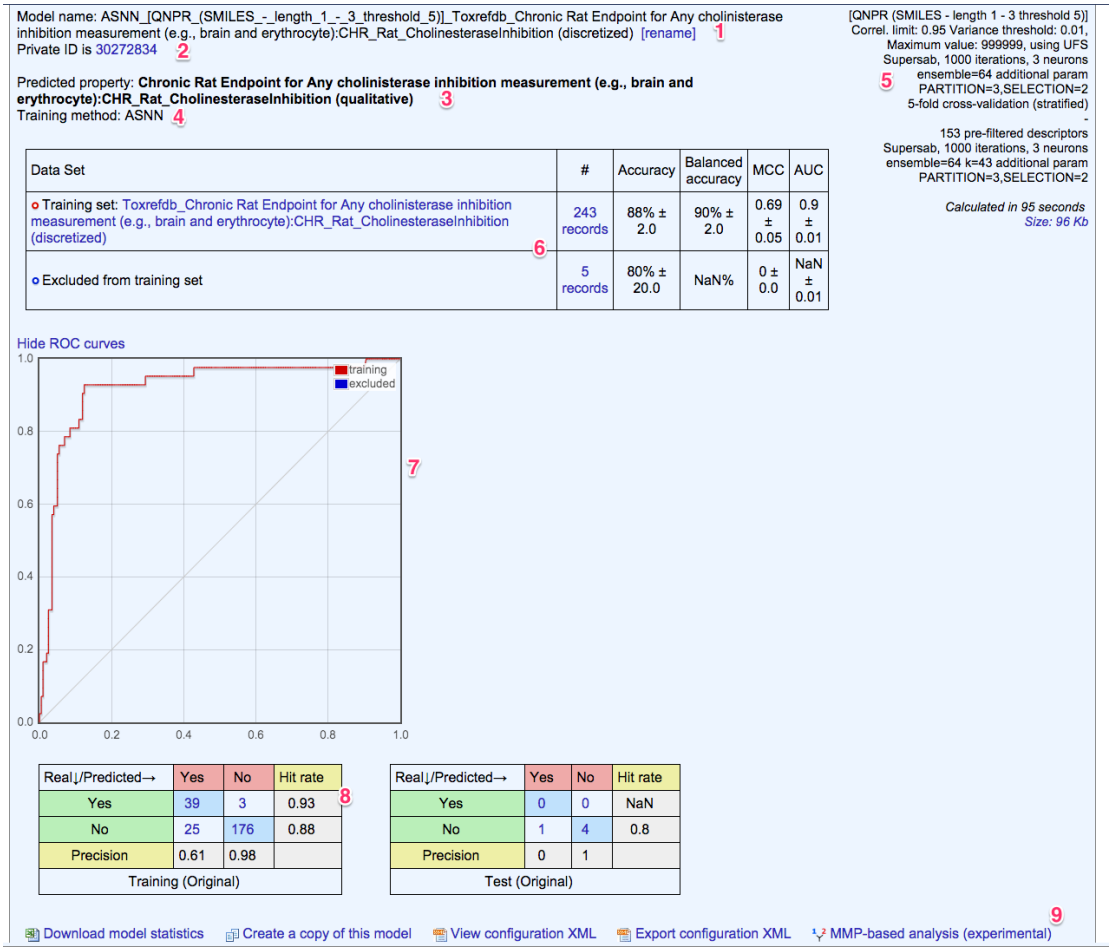


Figure 6

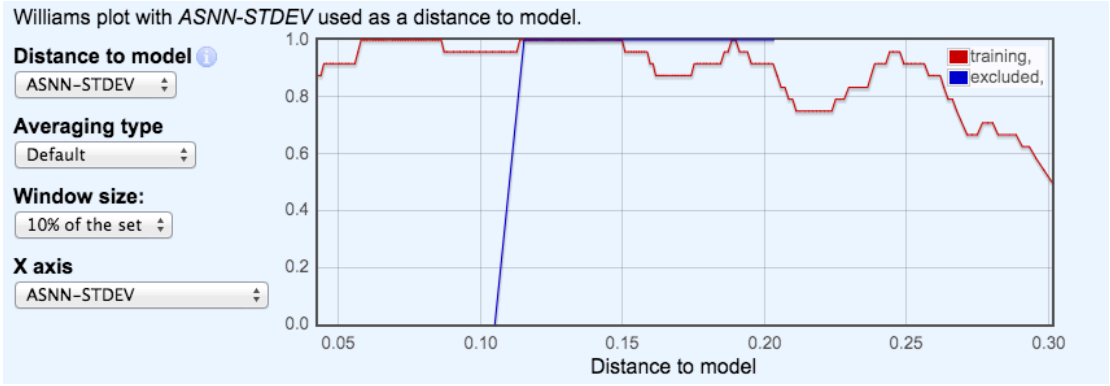


Figure 7

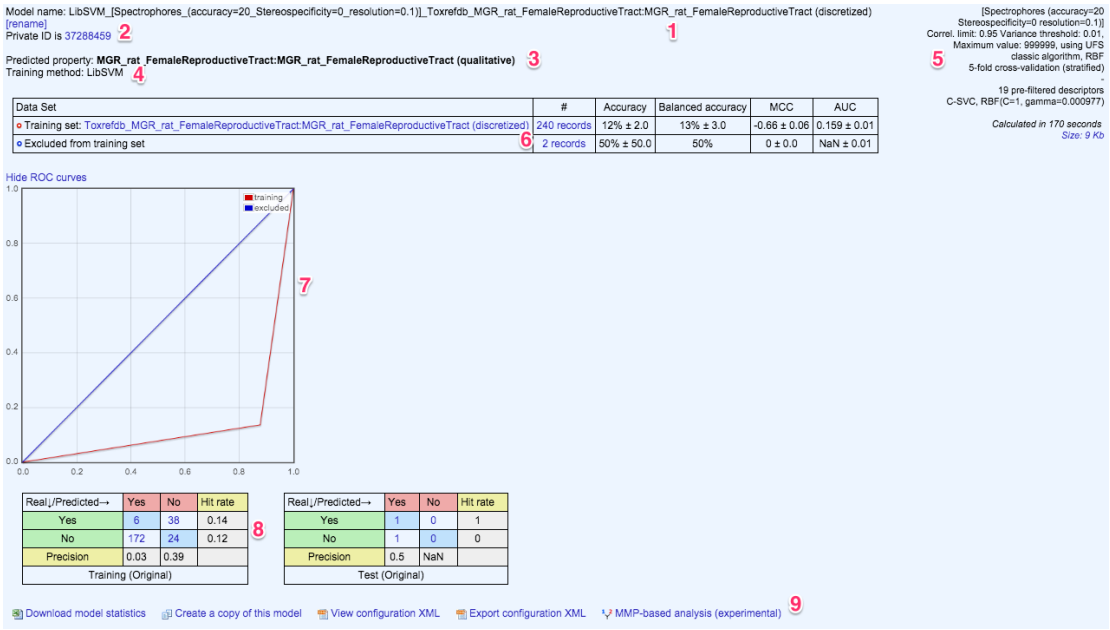


Figure 8

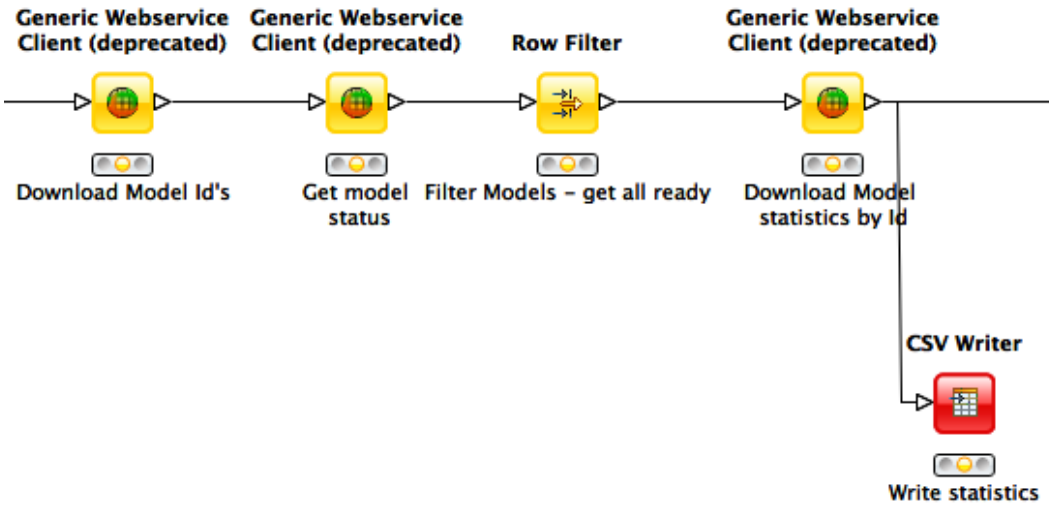


Figure 9

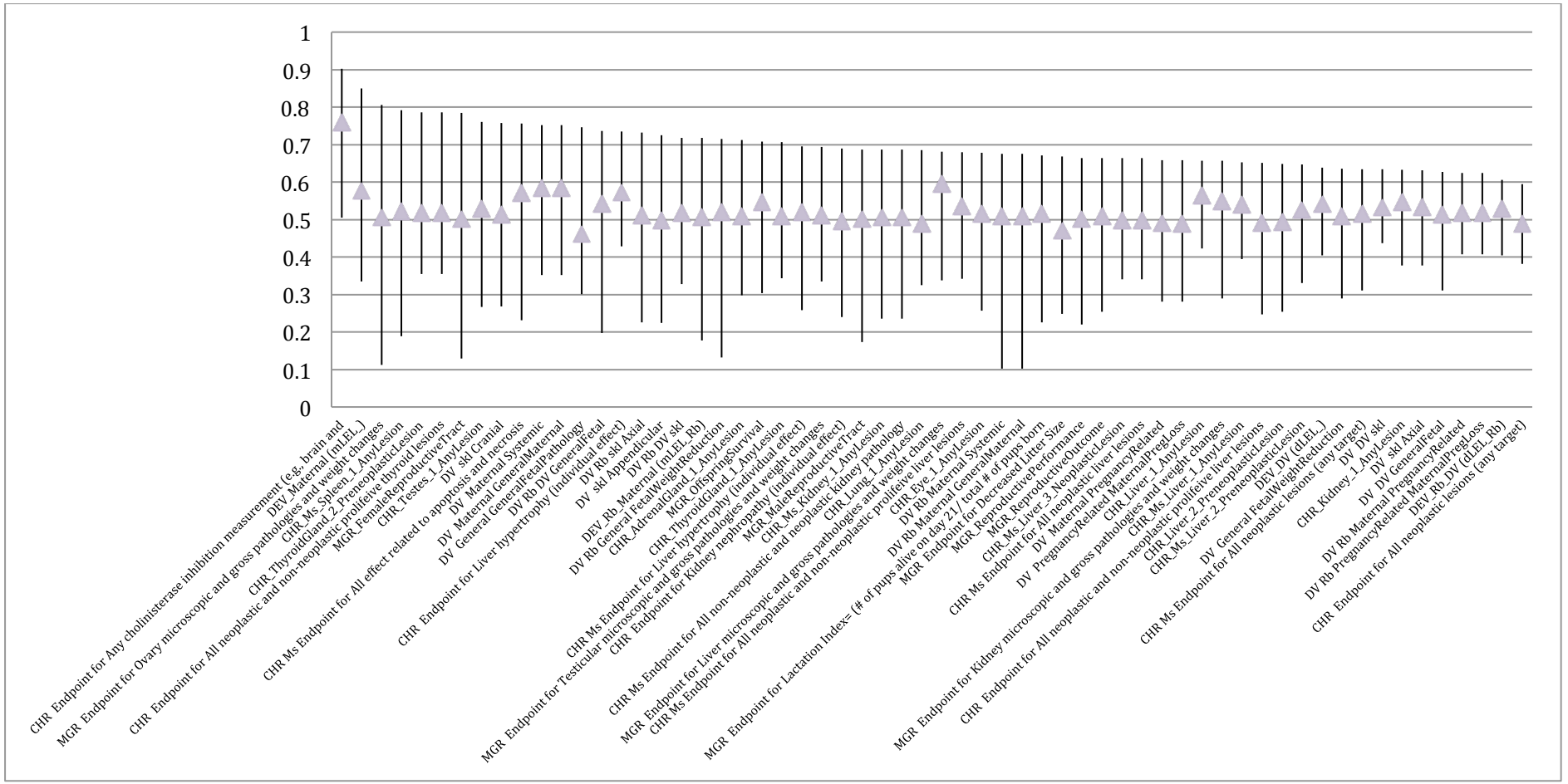

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
CHR_Endpoint for Any cholinisterase inhibition measurement (e.g., brain and
DEV_Maternal (mLEL_)
MGR_Endpoint for Ovary microscopic and gross pathologies and weight changes
CHR_Ms_Spleen_1_AnyLesion
CHR_ThyroidGland_2_PreneoplasticLesion
CHR_Endpoint for All neoplastic and non-neoplastic proliferative thyroid lesions
MGR_FemaleReproductiveTract
CHR_Testes_1_AnyLesion
DV_skl Cranial
DV_Maternal Systemic
DV_Maternal GeneralMaternal
DV_General GeneralFetalPathology
CHR_Endpoint for Liver hypertrophy (individual effect)
DV_Rb DV GeneralFetal
DV_Rb skl Axial
DV_skl Appendicular
DV_Rb DV skl
DEV_Rb_Maternal (mLEL_Rb)
DV_Rb General FetalWeightReduction
CHR_AdrenalGland_1_AnyLesion
MGR_OffspringSurvival
CHR_ThyroidGland_1_AnyLesion
CHR Ms Endpoint for Liver hypertrophy (individual effect)
MGR_Endpoint for Testicular microscopic and gross pathologies and weight changes
CHR_Endpoint for Kidney nephropathy (individual effect)
MGR_MaleReproductiveTract
CHR_Ms_Kidney_1_AnyLesion
CHR_Lung_1_AnyLesion
CHR_Eye_1_AnyLesion
DV_Rb Maternal GeneralMaternal
DV_Rb Maternal Systemic
MGR_Endpoint for Lactation Index= (# of pups alive on day 21/ total # of pups born
MGR Endpoint for Decreased Litter Size
MGR_ReproductiveOutcome
CHR Ms Endpoint for All neoplastic liver lesions
CHR_Ms_Liver_3_NeoplasticLesion
MGR_Endpoint for All neoplastic liver lesions
DV_Maternal PregnancyRelated
DV_PregnancyRelated MaternalPregLoss
CHR_Liver_1_AnyLesion
CHR_Ms_Liver_1_AnyLesion
MGR_Endpoint for Kidney microscopic and gross pathologies and weight changes
CHR_Ms_Liver_2_PreneoplasticLesion
CHR Ms Endpoint for All neoplastic and non-neoplastic proliferative liver lesions
DEV_DV (dLEL_)
DV_General FetalWeightReduction
CHR_Kidney_1_AnyLesion
DV_DV skl
DV_DV GeneralFetal
DV_skl Axial
DV_Rb Maternal PregnancyRelated
DV_Rb PregnancyRelated MaternalPregLoss
CHR_Endpoint for All neoplastic lesions (any target)
DEV_Rb_DV (dLEL_Rb)
CHR_Endpoint for All neoplastic lesions (any target)

Figure 10.

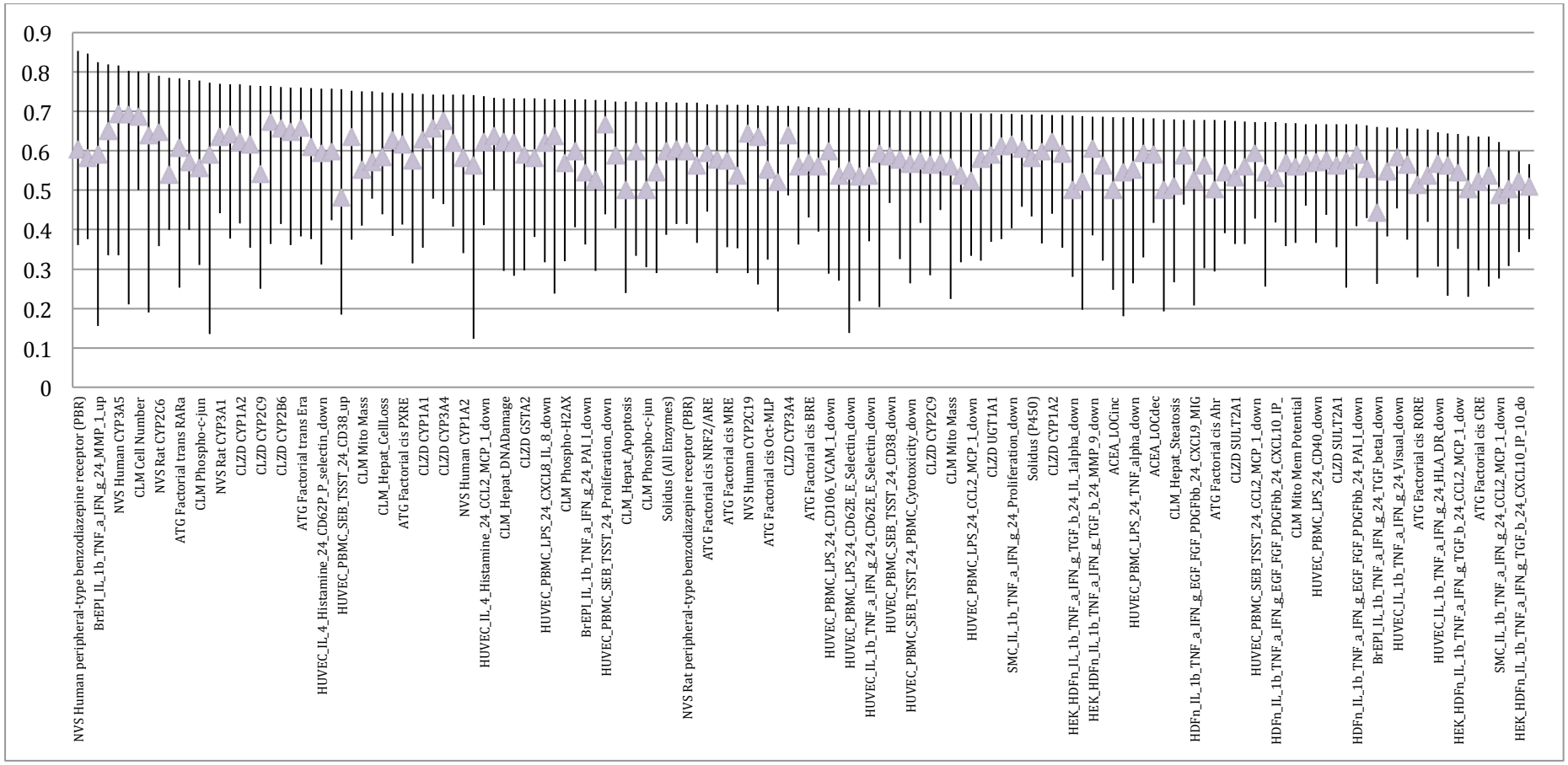

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
NVS Human peripheral-type benzodiazepine receptor (PBR)
BrEPI_IL_1b_TNF_a_IFN_g_24_MMP_1_up
NVS Human CYP3A5
CLM Cell Number
NVS Rat CYP2C6
ATG Factorial trans RARa
CLM Phospho-c-jun
NVS Rat CYP3A1
CLZD CYP1A2
CLZD CYP2C9
CLZD CYP2B6
ATG Factorial trans Era
HUVEC_IL_4_Histamine_24_CD62P_P_selectin_down
HUVEC_PBMC_SEB_TSST_24_CD38_up
CLM Mito Mass
CLM_Hepat_CellLoss
ATG Factorial cis PXRE
CLZD CYP1A1
CLZD CYP3A4
NVS Human CYP1A2
HUVEC_IL_4_Histamine_24_CCL2_MCP_1_down
CLM_Hepat_DNADamage
CLZD GSTA2
HUVEC_PBMC_LPS_24_CXCL8_IL_8_down
CLM Phospho-H2AX
BrEPI_IL_1b_TNF_a_IFN_g_24_PAI_1_down
HUVEC_PBMC_SEB_TSST_24_Proliferation_down
CLM_Hepat_Apoptosis
CLM Phospho-c-jun
Solidus (All Enzymes)
NVS Rat peripheral-type benzodiazepine receptor (PBR)
ATG Factorial cis NRF2/ARE
ATG Factorial cis MRE
NVS Human CYP2C19
ATG Factorial cis Oct-MLP
CLZD CYP3A4
ATG Factorial cis BRE
HUVEC_PBMC_LPS_24_CD106_VCAM_1_down
HUVEC_PBMC_LPS_24_CD62E_E_Selectin_down
HUVEC_IL_1b_TNF_a_IFN_g_24_CD62E_E_Selectin_down
HUVEC_PBMC_SEB_TSST_24_CD38_down
HUVEC_PBMC_SEB_TSST_24_PBMC_Cytotoxicity_down
CLZD CYP2C9
CLM Mito Mass
HUVEC_PBMC_LPS_24_CCL2_MCP_1_down
CLZD UGT1A1
SMC_IL_1b_TNF_a_IFN_g_24_Proliferation_down
Solidus (P450)
CLZD CYP1A2
HEK_HDFn_IL_1b_TNF_a_IFN_g_TGF_b_24_IL_1alpha_down
HEK_HDFn_IL_1b_TNF_a_IFN_g_TGF_b_24_MMP_9_down
ACEA_LOCinc
HUVEC_PBMC_LPS_24_TNF_alpha_down
ACEA_LOCdec
CLM_Hepat_Steatosis
HDFn_IL_1b_TNF_a_IFN_g_EGF_FGF_PDGFbb_24_CXCL9_MIG
ATG Factorial cis Ahr
CLZD SULT2A1
HUVEC_PBMC_SEB_TSST_24_CCL2_MCP_1_down
HDFn_IL_1b_TNF_a_IFN_g_EGF_FGF_PDGFbb_24_CXCL10_IP_
CLM Mito Mem Potential
HUVEC_PBMC_LPS_24_CD40_down
CLZD SULT2A1
HDFn_IL_1b_TNF_a_IFN_g_EGF_FGF_PDGFbb_24_PAI_1_down
BrEPI_IL_1b_TNF_a_IFN_g_24_TGF_beta1_down
HUVEC_IL_1b_TNF_a_IFN_g_24_Visual_down
ATG Factorial cis RORE
HUVEC_IL_1b_TNF_a_IFN_g_24_HLA_DR_down
HEK_HDFn_IL_1b_TNF_a_IFN_g_TGF_b_24_CCL2_MCP_1_dow
ATG Factorial cis CRE
SMC_IL_1b_TNF_a_IFN_g_24_CCL2_MCP_1_down
HEK_HDFn_IL_1b_TNF_a_IFN_g_TGF_b_24_CXCL10_IP_10_do

Figure 11

9 Tables

Table 1 List of *in silico* and biological descriptor packages used in the study. The number of descriptors within the package is shown.

Descriptors package name	Type	Descriptors count
AlogPS [53,54]+ Estate indices [55]	in silico	2+ 222
Chemaxon descriptors [56]	in silico (3D)	465
GSFragments [57]	in silico	588
ISIDA fragments [58]	in silico	1487
CDK [59]	in silico (3D)	204
Dragon 6 [60]	in silico (3D)	3127
inductive descriptors [61]	in silico (3D)	40
Mera + Mersy [62–64]	in silico (3D)	529 + 42
QNPR [65]	in silico	
Spectrophores [66,67]	in silico (3D)	144
Adriana.Code [68]	in silico (3D)	183
ToxCast *in vitro* assays	Biological	407
ToxCast *in vitro* assays + CDK	Biological + *in silico*	407 + 204
pathways perturbation	Biological	1178
pathways perturbation + CDK	Biological + *in silico*	1178 + 204
ToxCast *in vitro* assays + pathways perturbation	Biological	407 + 1178
ToxCast *in vitro* assays + pathways perturbation + CDK	Biological + *in silico*	407 + 1178 + 204

Table 2 Possible outcomes of a classification model. The table also lists the statistical parameters used for judging the quality of the QSAR models throughout the study. More statistical measures are presented in the supplementary materials.

	Experimental measurement	
Predicted outcome	True positive (TP)	False positive (FP)
	False negative (FN)	True negative (TN)
Sensitivity (SN) = TP / (TP + FN) Specificity (SP) = TN / (TN + FP) Accuracy (ACC) = (TP + TN) / (TP + FP + TN + FN) Balanced accuracy (BA) = (sensitivity + specificity) / 2 Matthews correlation coefficient (MCC) = (TP * TN – FP * FN) / [(TP + FP) (TN + FP) (TP + FN) (TN + FN)]$^{1/2}$		

Table 3. The five best predicted *in vitro* assays based on the maximum balanced accuracy of the respective models. SN: Sensitivity, SP: Specitificty, MCC: Matthews Correlation Coefficient, BA: balanced accuracy. Alg. Algorithm, Desc.: descriptor package. 68% confidence interval of the mean balanced accuracy is also shown. Models can be accessed from http://iprior.eadmet.com/model/[modelID]

Model id	Predicted endpoint	SN	SP	BA	MCC	Alg.	Desc..
30272834	Chronic Rat Endpoint for Any cholinisterase inhibition measurement (e.g., brain and erythrocyte)	0.93	0.88	0.90 ±0.03	0.69	ASNN	QNPR
47892767	DEV_rat_Maternal (mLEL_rat)	0.93	0.77	0.85 ±0.04	0.63	FSMLR	Inductive
17745510	Multigeneration Rat Endpoint for Ovary microscopic and gross	0.71	0.90	0.81±0.04	0.55	PLS	Chemaxon

	pathologies and weight changes						
23773279	CHR_Mouse_Spleen_1_AnyLesion	0.86	0.72	0.79±0.04	0.44	FSML R	ISIDA
9902926	CHR_Rat_ThyroidGland_2_PreneoplasticLesion	0.63	0.95	0.79±0.04	0.60	LibSVM	Dragon 6

Table 4 Most common toxicity alerts in the toxic acetylcholinesterase inhibitors showing clear indication of organphosphorus compounds

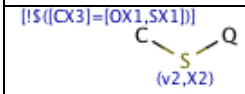
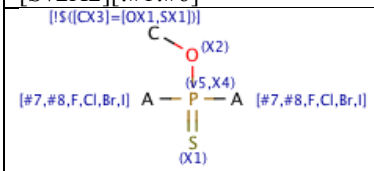
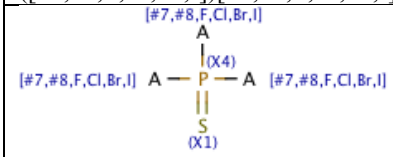
Toxicity Alert	# Toxic set (42)	# non-toxic set (206)	Enrichment factor	p-Value
 [#6&!$([CX3]=[OX1,SX1])][Sv2X2][!#1!#6]	14 (33.30%)	1 (0.50%)	68.7	10^{-11}
 [SX1]=[Pv5X4]([OX2][#6&!$([CX3]=[OX1,SX1])])([#7,#8,F,Cl,Br,I])[#7,#8,F,Cl,Br,I]	10 (23.80%)	1 (0.50%)	49	10^{-8}
 [SX1]=[Pv5X4]([#7,#8,F,Cl,Br,I])([#7,#8,F,Cl,Br,I])[#7,#8,F,Cl,Br,I]	10 (23.80%)	1 (0.50%)	49	10^{-8}

Table 5 Most significant *in vitro* assays in the toxic acetylcholinesterase inhibitors showing the impact on acetylcholinesterase enzyme

Toxicity Alert	# Toxic set (42)	# non-toxic set (206)	Enrichment factor	p-Value
ToxCast assay: Novascreen Rat AChE	13 (31.0%)	1 (0.5%)	63.8	10^{-10}
Toxcast_Pathway: 512 Glycerophospholipid metabolism	13 (31.0%)	1 (0.5%)	63.8	10^{-10}
Toxcast_Pathway: 801 Process: response to wounding GO id:0009611	10 (23.8%)	1 (0.5%)	49	10^{-8}
Toxcast_Pathway: 796 -- Component: basal lamina Description: Component: basal lamina accession: GO id:0005605	10 (23.8%)	1 (0.5%)	49	10^{-8}
Toxcast_Pathway: 793 Function: acetylcholinesterase activity GO id: 0003990	10 (23.8%)	1 (0.5%)	49	10^{-8}
Novascreen Human AChE	10 (23.8%)	1 (0.5%)	49	10^{-8}

Table 6 Toxicity endpoints where the biological descriptors contributed to the QSAR model with the highest balanced accuracy. The algorithm used for such model is shown (Alg.). Balanced accuracies (BA) for models developed using CDK (as an example for *in silico* descriptors) as well as different biological descriptors are

shown (68% confidence interval). Asys: ToxCast Assays, Pwys: ToxCast pathways perturbation. The id for the best model is shown for reference on iPrior website

Toxicity end point	Best model (out of 136 models per end point)			Balanced accuracies					
	id	BA	Alg.	CDK	Asys	Pwys	Asys+ pwys	CDK + Asys	CDK + pwys
MGR rat Female Reproductive Tract	28383900	0.78 ±0.04	FSMLR	0.59 ±0.04	0.61 ±0.04	0.62 ±0.04	0.57 ±0.04	0.79 ±0.04	0.6 ±0.04
Developmental Rat General Fetal Pathology	43415622	0.75 ±0.04	FSMLR	0.45 ±0.05	0.3 ±0.04	0.44 ±0.05	0.39 ±0.04	0.52 ±0.04	0.75 ±0.04
MGR rat Reproductive Performance	16914893	0.66 ±0.04	MLRA	0.49 ±0.04	0.57 ±0.04	0.56 ±0.04	0.66 ±0.04	0.51 ±0.04	0.42 ±0.04
Developmental Rat Pregnancy Related Maternal Preg Loss	6141411	0.66 ±0.03	FSMLR	0.46 ±0.03	0.66 ±0.03	0.52 ±0.03	0.63 ±0.03	0.55 ±0.02	0.52 ±0.03
Developmental rat Maternal Pregnancy Related	23386741	0.66 ±0.03	FSMLR	0.46 ±0.03	0.66 ±0.03	0.52 ±0.03	0.63 ±0.03	0.55 ±0.02	0.52 ±0.03
Chronic Rat Endpoint for All neoplastic and non-neoplastic proliferative liver lesions	28198946	0.65 ±0.04	FSMLR	0.44 ±0.03	0.42 ±0.04	0.65 ±0.04	0.49 ±0.03	0.55 ±0.03	0.55 ±0.03
Developmental Rat General Fetal Weight Reduction	2470006	0.64 ±0.03	LibSVM	0.63 ±0.03	0.5 ±0.03	0.58 ±0.03	0.64 ±0.03	0.47 ±0.03	0.53 ±0.03
Developmental rat Developmental Skeletal	13886224	0.63 ±0.03	WEKA-RF	0.58 ±0.03	0.47 ±0.03	0.53 ±0.03	0.55 ±0.03	0.53 ±0.03	0.63 ±0.03

Table 7. Comparing the performance of different descriptor packages in constructing QSAR models for *in vivo* toxicity and *in vitro* assays. For each descriptor package, the number of toxicity endpoints / *in vitro* assays where the descriptor package was able to contribute to the model with highest balanced accuracy is shown.

Descriptors	In vivo rank	# in vivo endpoints	In vitro rank	# in vitro endpoints
ALogPS+ OEstate	1	11	2	21
ISIDA Fragmentor	2	7	3	19
InductiveDescriptors	3	5	4	15
Spectrophores	4	5	10	5
CDK	5	4	5	15
ChemaxonDescriptors	6	4	1	22
Dragon6	7	4	9	9
GSFrag	8	4	7	12
QNPR	9	4	8	10
Adriana	10	3	6	14
Mera,_Mersy	11	2	11	2
CDK+ Toxcast_Pathways	12	2	NA	NA
Toxcast_assays	13	2	NA	NA
Toxcast_assays+	14	2	NA	NA

Toxcast_Pathways				
CDK+Toxcast_assays	15	1	NA	NA
Toxcast_Pathways	16	1	NA	NA

Table 8. Comparing the performance of different algorithms in constructing QSAR models for *in vivo* toxicity and *in vitro* assays. For each descriptor package, the number of toxicity endpoints / *in vitro* assays where the algorithm was able to contribute to the model with highest balanced accuracy is shown.

Algorithm	In vivo rank	# in vivo endpoints	In vitro rank	# in vitro endpoints
FSMLR	1	21	1	40
LibSVM	2	13	2	34
MLRA	3	7	6	9
KNN	4	6	7	9
WEKA-J48	5	6	4	14
ASNN	6	3	3	23
WEKA-RF	7	3	5	13
PLS	8	2	8	2

Table 9 The five best predicted *in vitro* assays based on the maximum balanced accuracy of the respective models. SN: Sensitivity, SP: Specitificty, MCC: Matthews Correlation Coefficient, BA: balanced accuracy. 68% confidence interval of the mean balanced accuracy is also shown

Model id	property	SN	SP	BA	MCC	Algorithm	Descriptors
1659347	Novascreen Human peripheral type benzodiazepine receptor (PBR)	0.83	0.87	0.85 ±0.03	0.59	FSMLR	Inductive Descriptors
32112993	HUVEC_IL_1b TNF_a_IFN_g_24_SRB_down	0.72	0.97	0.85 ±0.03	0.74	LibSVM	Spectrophores
10535027	BrEPI_IL_1b_TNF_a_IFN_g_24_MMP_1_up	0.71	0.94	0.83 ±0.04	0.62	LibSVM	ALogPS, OEstate
30522735	Novascreen Human CYP2B6	0.78	0.86	0.82 ±0.04	0.50	FSMLR	ALogPS, OEstate
44227184	Novascreen Human CYP3A5	0.87	0.76	0.82 ±0.03	0.49	LibSVM	Dragon6