Published in final edited form as:
Chem Res Toxicol. 2017 November 20; 30(11): 2046–2059. doi:10.1021/acs.chemrestox.7b00084.

Predicting organ toxicity using *in vitro* bioactivity data and chemical structure

Jie Liu[2,3], **Grace Patlewicz**[1], **Antony J. Williams**[1], **Russell S. Thomas**[1], and **Imran Shah**[1,*]

[1]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, 27711, USA

[2]Department of Information Science, University of Arkansas at Little Rock, AR, 72204, USA

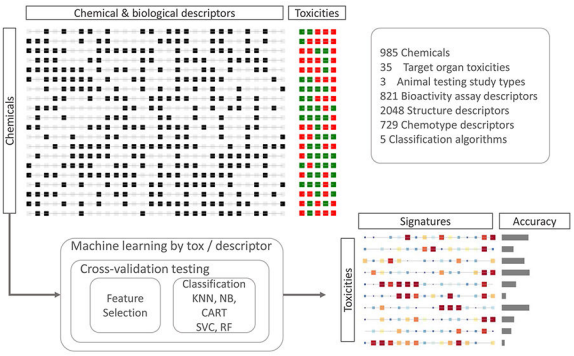
[3]Oak Ridge Institute for Science Education Fellow, National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, 27711, USA

Abstract

Animal testing alone cannot practically evaluate the health hazard posed by tens of thousands of environmental chemicals. Computational approaches making use of high-throughput experimental data may provide more efficient means to predict chemical toxicity. Here, we use a supervised machine learning strategy to systematically investigate the relative importance of study type, machine learning algorithm, and type of descriptor on predicting *in vivo* repeat-dose toxicity at the organ-level. A total of 985 compounds were represented using chemical structural descriptors, ToxPrint chemotype descriptors, and bioactivity descriptors from ToxCast *in vitro* high-throughput screening assays. Using ToxRefDB, a total of 35 target organ outcomes were identified that contained at least 100 chemicals (50 positive and 50 negative). Supervised machine learning was performed using Naïve Bayes, k-nearest neighbor, random forest, classification and regression trees, and support vector classification approaches. Model performance was assessed based on F1 scores using five-fold cross-validation with balanced bootstrap replicates. Fixed effects modeling showed the variance in F1 scores was explained mostly by target organ outcome, followed by descriptor type, machine learning algorithm, and interactions between these three factors. A combination of bioactivity and chemical structure or chemotype descriptors were the most predictive. Model performance improved with more chemicals (up to a maximum of 24%) and these gains were correlated ($\rho = 0.92$) with the number of chemicals. Overall, the results demonstrate that a combination of bioactivity and chemical descriptors can accurately predict a range of target organ toxicity outcomes in repeat-dose studies, but specific experimental and methodologic improvements may increase predictivity.

Abstract

*Corresponding author information: Imran Shah, Tel: (919) 541 1391 shah.imran@epa.gov.



Keywords

machine learning; prediction; (Q)SAR; bioactivity; chemotypes; ToxCast; ToxRefDB

1 INTRODUCTION

The US Environmental Protection Agency (EPA), other international regulatory agencies, and pharmaceutical and consumer product companies need new tools to efficiently and effectively assess toxicity across a large number of chemicals. Under the current system, determining the toxicological hazards posed by chemicals relies heavily on animal experimentation.[1] In a regulatory risk assessment context, these experiments are typically based on test guidelines whose protocols have been standardized and agreed upon across international agencies (e.g., OECD). Depending on the toxicological endpoint of interest and the specific test guideline, a single study can take a year or more to complete and costs thousands to more than a million dollars.[2] In addition, multiple test guideline studies are often required to fully evaluate a chemical for potential toxicities.

There are more than 85,000 substances[3] listed on the Toxic Substances Control Act (TSCA 1976) inventory, and close to 144,000 chemicals[4] pre-registered under the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH 2006) regulation. The cost and duration of animal testing render it virtually impossible to fully evaluate the health hazards posed by the many thousands of environmental and industrial chemicals on these lists[5]. A long-term strategy for addressing this challenge was presented in the National Research Council (NRC) report entitled, *Toxicity Testing in the 21st Century: A Vision and a Strategy*.[6] This vision of 21st-century toxicology called for a change in toxicity testing to utilize *in vitro* high-throughput screening (HTS) assays to evaluate thousands of chemicals for their molecular effects on specific pathways linked to toxicity. The US EPA, National Institutes of Health, and National Toxicology Program formed the Tox21 partnership[7] to implement the NRC vision. In addition, the US EPA undertook implementation of the NRC vision through the ToxCast project. [8,9,10] Collectively, Tox21 and ToxCast have generated one of the largest data sets on biological activities as related to environmental and industrial chemicals.[11]

A key objective of ToxCast and Tox21 has been to use *in vitro* bioactivity data to identify potential hazards and prioritize data-poor chemicals for additional testing and assessment.[9]

The identification of potential hazards and prioritization for additional testing has occurred using a variety of modeling and analysis approaches. First, the *in vitro* bioactivity data have been used to identify chemicals that activate molecular initiating events (MIEs) in adverse outcome pathways (AOPs).[12] One example of a model of an MIE is thyroperoxidase inhibition for chemically-induced thyroid toxicity.[13] Second, *in vitro* bioactivity assays from ToxCast have been aggregated in a biological pathway or process context to prioritize chemicals for additional study or predict toxicity including endpoints such as endocrine disruption,[14] obesity,[15] development, [16] and cancers.[17] Finally, *in vitro* bioactivity has been used in supervised machine learning analyses to predict a range of toxicological responses from *in vivo* studies including developmental,[18] reproductive[19] and chronic toxicity.[20]

Machine learning methods are also widely used for building quantitative structure–activity relationships (QSAR). QSARs are classification and regression models for mapping molecular structural features of chemicals to their physical, chemical, or biological properties (see https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology). While a complete review of QSAR models is beyond the scope of this work, the work of Low *et al*[21] on integrating chemical and bioactivity (genomics) data for predicting hepatotoxicity is relevant for introducing the notion of "hybrid" descriptors and predictive models. Motivated by their work, we recently used a combination of chemical structure and *in vitro* bioactivity to predict histopathological sub-classes of rodent hepatotoxicity using hybrid models.[22]

One of the most important challenges in building useful predictive models for evaluating chemical safety is identifying high quality chemistry, bioactivity and toxicology data.[23] We recognize the importance of curation[24],[25] in building publicly available chemistry resources. Each of the sources of data used in this analysis, including chemical[11], bioactivity[26] and toxicity[27] data are continuously evaluated for quality and updates are released publicly.

In this study, we expanded on the approaches used to predict hepatotoxicity[22] and used a similar supervised machine learning strategy to predict 35 *in vivo* target organ toxicity outcomes across a range of repeat-dose guideline study types. We objectively evaluated three main types of chemical descriptors including 821 *in vitro* HTS assay endpoints from ToxCast, 2,048 extended connectivity chemical fingerprints (specifically Morgan fingerprints), and 729 expert-derived chemical descriptors (namely ToxPrint chemotypes). We also evaluated the performance of two types of hybrid bioactivity and chemical structure descriptors by combining the *in vitro* HTS assays endpoints with Morgan fingerprints and with chemotypes. We treated each target organ toxicity as a separate class and used five-fold cross-validation testing to systematically analyze factors that impact classification performance. The main factors include five different descriptor types, eight learning algorithms, and the number of descriptors. Our analysis provides an estimate of the best baseline predictive accuracy by study and target organ toxicity.

2 MATERIALS AND METHODS

1. Overview of the approach

Our analysis comprised environmental and pharmaceutical chemicals from the ToxCast and ToxRefDB data sets. The ToxCast data set included *in vitro* bioactivity measurements from 821 HTS assays, while ToxRefDB incorporates data from up to 10 different *in vivo* guideline testing study types. Across the different *in vivo* guideline studies, we identified 35 target-organ toxicity outcomes across four study types containing at least 100 chemicals (50 positives and 50 negative chemicals) and used a supervised machine learning approach to comprehensively examine the ability of chemical structure and *in vitro* bioactivity to predict the potential for target organ toxicity.

2. Data sources

***In vivo* animal toxicity data—**Toxicity data were obtained from ToxRefDB,[27] which describes the *in vivo* effects of hundreds of compounds from animal testing studies. All data are publicly available for download (https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data; Version Nov2014). ToxRefDB includes information about the effects of chemicals on different species and target organs in US EPA, OECD or other regulatory guideline or guideline-like studies. For this analysis, *in vivo* outcomes for 985 chemicals taken from ToxRefDB were aggregated at the level of a study type and target site of the effects (e.g., the target organ affected) across multiple species. The study types included chronic toxicity (CHR), developmental toxicity (DEV), developmental neurotoxicity (DNT), multigenerational toxicity (MGR), neurotoxicity (NEU), reproductive toxicity (REP), acute toxicity (ACU), sub-acute toxicity (SAC), and sub-chronic toxicity (SUB). All other toxicity testing studies (i.e., where a specific guideline was not reported) were grouped into a category that was referred to as “other” (OTH). There were 129 unique target effects in ToxRefDB (found in the column “effect_target” of the ToxRefDB data file). If a chemical produced a significant effect in a particular study type at any dose, and in any species, then it was categorized as positive for that target effect. Effects not associated with a specific target organ such as body weight changes, clinical chemistry, and urinalysis, were excluded from the analysis. For example, if a chemical produced statistically significant treatment-related effects and, therefore, resulted in an assignment of a “positive” call for liver and kidney effects in a sub-chronic study, then it was considered both a “sub-chronic liver toxicant” (denoted as, SUB:Liver), and “sub-chronic kidney toxicant” (denoted as, SUB:Kidney). If a substance did not produce any statistically significant treatment-related effect on a target site whose evaluation was required in a study, then a negative call was assigned. The number of negative chemicals for all organs was inferred because their evaluation is mandatory in these studies. For example, the assessment of hepatic effects is necessary for guideline chronic testing studies. As a result, chemicals that were tested in chronic studies but do not produce hepatic changes were considered negative for chronic liver toxicity. However, it is not possible to infer negative chemicals for organs whose assessment is not compulsory in a guideline study. In these cases we could not distinguish between untested and negative chemicals. For instance, the evaluation of hepatic effects is not a requirement in multigenerational and developmental toxicity studies. As a result, we

know the number of positive chemicals but not the number of negative chemicals for hepatic effects.

Negative or positive effects were aggregated across sex and species within a study type but not across study types (i.e., a chemical could have a positive effect in a chronic study, but be negative in a subchronic study). We recognize that summarizing different effects across different types of clinical outcomes in different species oversimplifies complex pathology. However, in the interest of investigating whether any inferences could be made and quantified for a large number of substances, this type of aggregation was a necessary, yet pragmatic, approach to performing the subsequent computational analysis. All toxicity data are provided as supplemental material (S1).

***In vitro* bioactivity data—**The *in vitro* assay data were generated from the *in vitro* HTS of ToxCast Phase I and Phase II compounds and is publicly available (https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data; version Nov2014). The data were collected on 821 HTS assays from 6 different technology platforms: ACEA Biosciences (ACEA); Apredica (APR); Attagene, Inc. (ATG); NovaScreen panel (NVS); Odyssey Thera (OT); and Tox21. In ToxCast, each assay datum was reported as the chemical concentration (micromolar) at half maximal efficacy (AC50). An overall activity call is also made based on whether there is a statistically significant concentration-response. The criteria for determining a statistically significant response varies with the type of assay but is generally a multiple of the baseline median absolute deviation or minimal efficacy cutoff. Full documentation regarding the criteria for determining a significant response can be found on the ToxCast data download page (https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data).

For the machine learning analysis, all active and inactive assay results were converted to binary values (active=1 and inactive=0, respectively). The bioactivity data (denoted as, X^{bio}) of each chemical (i) is given by: $X_i^{bio} = \left\{x_i^1, x_i^2, x_i^3, \ldots, x_i^{bio}, \ldots, x_i^{n^{bio}}\right\}$, where x_i^{bio} represents a bioactivity assay (*bio*) result. When more than one bioactivity value was available for a given chemical-assay pair, the results were combined with a logical OR operation (that is, $x^{bio} = x_i^{bio} \vee x_j^{bio}$). The total number of bioactivity assays are denoted as n^{bio} (n^{bio} =821). All bioactivity (denoted as, bio) data are available as supplemental material (S2).

Each of the bioactivity descriptors is based on a specific ToxCast HTS assay in which chemicals were tested. The assay designation includes some information about the technology, the biological target, the exposure duration, and the direction of the effect. For instance, assays beginning with ATG (Attagene) measure multiplexed transcriptional activities in human HepG2 hepatoma cells [28]. The “ATG NF kB CIS up” assay measures increase in *cis*-activation of the Rel protein, which is the DNA binding subunit of the NF-kappa-B (NFKB1) complex. Assays beginning with the prefix NVS (NovaScreen) measure protein activities in a cell-free system [29]. For instance, the “NVS NR hPPARg” assay measures the biochemical function of the human peroxisome proliferator activator receptor gamma (PPARG). Finally, the assays beginning with APR (Apredica) are conducted in human HepG2 hepatoma cells, and they combine fluorescently labeled antibodies with

automated imaging (high-content imaging) to measure cell level changes [30]. Some examples of the APR assays include: cJun phosphorylation at 1 h (denoted as, APR HepG2 StressKinase 1h up), increase in phosphorylated tubulin at 72 h (APR HepG2 MicrotubuleCSK 72h up), cell cycle progression changes at 72 h (APR HepG2 CellCycleArrest 72h dn) and a decrease in cell number at 24 h (APR HepG2 CellLoss 24h dn). Detailed descriptions of all assays are available from the ToxCast download page (https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data).

Chemical structure data—The chemical descriptors (chm) used in this study relied upon previously published structural fingerprints. Fingerprints were calculated using "QSAR-ready" chemical structures, obtained according to the chemical curation workflow reported by Mansouri *et al*[31], were obtained from the DSSTox[11] database. The workflow results in the removal of salts and inorganic counterions, the conversion of tautomers to unique representations, the neutralization of charged structures, when possible, and the removal of stereochemistry information. Fingerprints were represented as binary (bit) vectors where the elements themselves represent the presence or absence of a certain feature. We calculated structural fingerprints using Morgan fingerprints,[32] which are a type of extended-connectivity fingerprint. They are constructed by defining the molecular subgraphs in the neighborhood of each non-hydrogen atom in circular layers up to a defined diameter. These subgraphs are mapped to integer codes by a hashing procedure resulting in a bit vector. The Morgan fingerprints were calculated using the freely available python RDKit cheminformatics library.[33] The total number of chemical descriptors were denoted as n^{chm} (n^{chm} = 2,048). The fingerprints were generated for 1,733 chemicals, and all chemical structure data are provided as supplemental material (S3).

ToxPrint Chemotype data—ToxPrint chemotypes are publicly available, expert-derived structural fragments based on medicinal chemistry.[34] The ChemoTyper software was used to search the occurrence of 729 chemotypes in each chemical structure. The chemotype descriptors (ct) for each chemical were represented as a bit vector where presence was signified by 1 and absence by 0. Chemotype descriptors for all chemicals are provided as supplemental material (S4).

Hybrid descriptors—In addition to the three primary descriptor subtypes (i.e., bio, chm, and ct), we constructed two sets of "hybrid" descriptors for each compound. Hybrid descriptors represent high dimensional spaces formed by the union of disparate descriptor types.[21] First, we merged the *in vitro* bioactivity (bio) and chemical (chm) descriptors to create a set of hybrid descriptors (denoted as bc). Second, we merged *in vitro* bioactivity (bio) and chemotype (ct) descriptors to create a second set of hybrid descriptors (denoted as bct).

Toxicity data selection—A total of 35 target organ endpoints possessed a minimum subset of at least 50 positive and 50 negative chemicals. Each outcome was denoted by concatenating the study and target organ and denoted as β. For example, "CHR:Liver" and "MGR:Ovary" indicate chronic liver effects and multigenerational ovarian effects respectively. We constructed toxicity data sets (X) for the 35 target organ outcomes for each

of the five descriptor types including chemical (X^{chm}), bioactivity (X^{bio}), chemotype (X^{ct}), bioactivity and chemical (X^{bc}), and bioactivity and chemotype (X^{bct}).

Most of the target organ toxicity endpoints had more negative than positive chemicals (CHR:Liver, SUB:Liver and CHR:Kidney being the only exceptions). A significant imbalance in positive and negative examples in the training set can bias supervised machine learning algorithms and reduce their performance on unseen data. To avoid this bias we randomly sampled ten balanced subsets from each target organ toxicity data set (without replacement) beginning with 50 positive and negative chemicals up to the maximum number possible (in steps of ten positives and ten negatives). For each target organ toxicity endpoint the minimal dataset was defined as one with the smallest number of examples (i.e., 50 positives and 50 negatives). Similarly, we identified the full data set as one with the largest number of cases. We represented each balanced toxicity data subset (denoted as, $X_j^{\alpha,\beta}$) using the five different descriptors (denoted as, α, where $\alpha \in$ {chm, bio, ct, bc, bct}), and excluded chemicals with missing values for any of the descriptors. For example, the CHR:Kidney data set contained 539 chemicals with target organ toxicity data from ToxRefDB. Out of these 539 chemicals there were 324 positives and 215 negatives, and 7,056 descriptors in all. For these 539 chemicals there were: 532 chemicals with 1,992 chm descriptors, 428 chemicals with 777 bio descriptors, 421 chemicals with 370 ct descriptors, 421 chemicals with 1,147 bct descriptors, and 421 chemicals with 2,769 bc descriptors. We randomly sampled 10 balanced subsets for each descriptor type using 100, 120, 140, 160, 180, 200, 220, 240, 260, 280, 300, 320, 340, 360 and 380 chemicals (producing 15*10*5 = 750 data subsets). We applied the same procedure to create data subsets for each of the 35 target organ toxicities.

3. Supervised machine learning

Predictive models of each of the 35 target organ toxicity endpoints were developed using the chm, bio, ct, bc and bct descriptors via supervised machine learning. For each of the 35 endpoints balanced subsets containing equal numbers of positive and negative chemicals were constructed for cross-validation testing.

Classification algorithms—Five different classification algorithms were used including naïve Bayes (NB), support vector machines classification (SVC), classification and regression trees (CART), k-nearest neighbors (KNN), and random forest (RF). The NB algorithm is a probabilistic method based on Bayes theorem that assumes that all features are independent.[35] SVC approaches find decision boundaries that can maximize the margin between the positive and negative classes of compounds.[36, 37] The SVC were trained using both linear or radial basis kernels and the corresponding classifiers denoted as SVCL and SVCR respectively. CART-based methods partition the feature space into a set of rectangles and then fit a simple model in each one.[38] We trained the CART algorithm by either pruning decision trees to a maximum depth of 10 (denoted as, CART0), or by allowing the learning algorithm to find the best number of splits based on the data (denoted as, CART1). The KNN method assigns the label of its nearest neighbor to an observation and determines the class by majority vote.[35] The KNN classifiers were trained using k=3 nearest neighbors (KNN0) and k=5 nearest neighbors (KNN1).

Cross-validation testing—Five-fold cross-validation testing was used to evaluate model performance. Each data set was randomly partitioned into five equal size subsets. Four subsets were used for training while the last subset was used for testing. The cross-validation testing was repeated ten times. For each step in the cross-validation loop the subset of the best n_{ds} descriptors was filtered using the ANOVA F-value to measure the association between the class and the descriptor. To avoid over-fitting, the maximum number of descriptors was limited to 25 ($n_{ds} \leq 25$). Classification accuracy was evaluated on each iteration of the 5-fold cross-validation testing from which the mean and standard deviation were calculated for the different performance metrics. The F1 score is the weighted average of sensitivity and specificity defined as $\mathrm{F1} = \frac{2 * (sensitivity * specificity)}{(sensitivity + specificity)}$. Each toxicity data subset was analyzed ten times by five-fold cross-validation testing with in-loop descriptor selection. We selected the best 5 to 25 (with a step size of 1) descriptors using the ANOVA F-value. We recorded cross-validation performance using sensitivity, specificity, accuracy, and F1 score (along with means and standard deviations for each statistic). We also stored the entire toxicity data subset (defined by positive and negative chemicals) along with the performance results in a database.

The outline of the entire supervised machine learning workflow is given below:

1. Select target organ toxicity

Identify target organ toxicities (β) from ToxRefDB

1.1. Identify chemicals associated with β

Find the positive chemicals (I_β^+), negative chemicals (I_β^-), and total chemicals ($I_\beta = I_\beta^+ \, UI_\beta^-$)

$n_\beta^+ = \mathrm{number}(I_\beta^+)$, $n_\beta^- = \mathrm{number}(I_\beta^-)$ and $n_\beta = n_\beta^+ + n_\beta^-$

Identify β where $n_\beta^+ \geq 50$ and $n_\beta^- \geq 50$

Construct ($n_\beta \times 1$) binary vector X^β to represent positive (1) and negative (0) chemicals

1.2. Obtain data for chemicals associated with β

For descriptor type $\alpha \in \{\mathrm{chm}, \mathrm{bio}, \mathrm{ct}, \mathrm{bc}, \mathrm{bct}\}$ construct data matrices X^α for I_β chemicals

Construct $X^{\alpha,\beta}$ by merging X^α and X^β using unique chemical identifiers in I_β

2. Predict and evaluate toxicity using supervised machine learning

```
For each descriptor α ∈ {chm, bio, ct, bc, bct}:

    For n_i = 50 to min(n_β^+, n_β^- with stepsize of 10:

        Construct balanced subsets of X^α,β
```

Repeat 10 times:

I_i^+, I_i^- = random subset of n_i chemicals from I_β^+, I_β^- and $I_i = I_i^+ \cup I_i^-$

$X_j^{\alpha,\beta} = X^{\alpha,\beta}[I_i]$ i.e. $X_j^{\alpha,\beta} \subset X^{\alpha,\beta}$

Vary number of descriptors (n_{ds}) from 5 to 25:

Repeat 10 times:

Conduct 5-fold cross-validation testing:

Split $X_j^{\alpha,\beta}$ into $\{X_{j,1}^{\alpha,\beta}, X_{j,2}^{\alpha,\beta}, X_{j,3}^{\alpha,\beta}, X_{j,4}^{\alpha,\beta}, X_{j,5}^{\alpha,\beta}\}$ balanced subsets

For $k \in \{1,2,3,4,5\}$:

$X_{train} = X_{j,m}^{\alpha,\beta} \cup X_{j,m'}^{\alpha,\beta} \cup X_{j,m''}^{\alpha,\beta} \cup X_{j,m'''}^{\alpha,\beta}$ where $k \notin \{m,..,m'''\}$

$X_{test} = X_{j,k}^{\alpha,\beta}$

Build Classifier (C) using X_{train} with top n_{ds} descriptors using

ANOVA F-value

Test C using X_{test}

Save the performance scores for $\{\beta, \alpha, n_i, n_{ds}, C\}$

Statistical comparisons of classification methods—We used analysis of variance (ANOVA) to evaluate statistical differences in F1 scores between target organ toxicities, machine learning algorithms and descriptor types. First, we used one-way ANOVA to independently compare the impact of machine learning algorithm and descriptor type on F1 scores. This was followed by Tukey's honest significance difference (HSD) test for multiple comparisons of differences between means.[39] Second, we used fixed effects linear models to explain the variance in F1 scores using organ toxicity, machine learning algorithm, and descriptor type as well as interactions between these factors. We calculated the proportion of variance in F1 scores explained by each of the factors using the η^2 (Eta-squared) statistic.[40]

Identifying descriptors frequently used in classification models—We used the following approach to calculate the relative usage of each descriptor of type α (where $\alpha \in$ {chm, bio, ct, bc, bct}) across all machine learning models for predicting a target organ

toxicity β. First, accurate machine learning models (F1 score ≥ $F1_0$) constructed using different numbers of descriptors ($n_{ds} \leq n_{ds0}$) were identified. Second, the frequency of occurrence of each descriptor (x_i^α) across the machine learning models for each target organ toxicity outcome β was calculated (denoted as $f_{\beta,i}^\alpha$). Third, a descriptor-toxicity frequency matrix (denoted as Φ) was constructed in which the rows and columns corresponded to target organ toxicity outcomes (β) and descriptors x_i^α, respectively. This matrix was populated with frequencies $f_{\beta,i}^\alpha$, μ_β^α σ_β^α and the columns were sorted by their median value (to aid selection of most frequent descriptors). Finally, Φ was row standardized to correct for any differences in the number of models across target organ toxicities. Row standardization was carried out with the formula: $f_{\beta,i}^\alpha - \mu_\beta^\alpha \Big/ \sigma_\beta^\alpha$, where μ_β^α and σ_β^α are the row-wise means and standard deviations, respectively. Each row of the matrix (denoted as Φ_β) represents the descriptor "signature" for toxicity outcome β. In its simplest form, a signature Φ_β defines relevant descriptors ($\{x_1^\alpha, x_2^\alpha, \ldots, x_i^\alpha, \ldots\}$) for predicting a target organ toxicity outcome β along with scores ($\{\varphi_{\beta 1}^\alpha, \varphi_{\beta 2}^\alpha, \ldots, \varphi_{\beta i}^\alpha, \ldots\}$) to capture their relevance.

Software—Data processing and analysis was conducted in the Python programming language (version 2.7) using RDKit (version 2014-09-02),[33] and the matplotlib package (version 0.99.1.2)[41] for visualization. All code will be made available on GitHub under (github.com/i-shah/ml-organ-tox).

3 RESULTS

1. Data sets

A total of 47 target organs were identified with outcomes from five or more chemicals across at least one guideline study type (Figure 1). The five most frequent chronic organ outcomes in descending order were the liver (414 positives and 125 negatives), kidney (324 positives and 215 negatives), spleen (205 positives and 334 negatives), adrenal gland (188 positives and 351 negatives), and lung (183 positives and 356 negatives). These five organs were also frequent sites of chemical effects in subchronic, and multigenerational studies. A decrease in the number of positive chemicals for these organs was observed in the subchronic, multigenerational and developmental studies when compared to the chronic studies. With the exception of the liver and kidney endpoints in the chronic and subchronic study types, more negative than positive chemicals were identified. Across all endpoints and study types, the negative chemicals outweighed the positive chemicals by a 3.4:1 ratio (on average). From the larger set of target organ endpoints and study types, 35 outcomes possessed at least 50 positive and 50 negative chemicals that could be used in the supervised machine learning analysis. These 35 outcomes comprised of 20 target organs and three guideline study types. The target organs included: adrenal gland, bone marrow, brain, eye, heart, kidneys, liver, lungs, lymph nodes, mammary glands, ovaries, pancreas, pituitary gland, spleen, stomach, testes, thymus, thyroid gland, urinary bladder, and uterus. In descending order of the number of outcomes, chronic studies showed the greatest number of chemical effects (19/35; 54.3%)

followed by subchronic (12/35; 34.3%), and multigenerational (4/35; 11.4%). None of the 35 outcomes were observed in developmental studies.

2. Predictive accuracy using minimal data sets

To establish the performance baseline for predicting target organ toxicity, we used the minimal data sets (defined by 50 positive and 50 negative chemicals) for each of the 35 target organ outcome and study type pairs. Across all toxicities, descriptor types and classifiers, the maximum F1 score was 0.85±0.09 for predicting MGR:Brain) (Figure. 2). There was no appreciable increase in the F1 score beyond 24 descriptors for most target organ toxicities except for chronic liver and kidney outcomes.

A broad assessment of cross-validation performance was undertaken for the minimal data set across all machine learning algorithms, descriptor types, and target organ outcomes (Figure 3; Supplemental Material, S5). The mean F1 score across all target organ outcomes, machine learning algorithms, and descriptor types was 0.69. When broken down by machine learning algorithm, the mean F1 scores across all target organ data sets and descriptor types in descending order were: KNN0 0.73, KNN1 0.72, RF0 0.70, SVCR0 0.70, CART 0.69, SVCL0 0.69 and NB 0.61. Based on Tukey's HSD post hoc analysis, differences in mean F1 scores between machine learning algorithms were statistically significant ($p<0.01$) except between CART0 and CART1, and between KNN0 and KNN1. NB algorithms produced classification models with the greatest variability across all performance metrics. Notably, NB models together with chm descriptors consistently produced the least accurate classification models. In contrast, KNN-based classification models were generally the most sensitive, but also the least specific across all target organ toxicities. Decision tree-based algorithms (RF and CART) and support vector classification models were distinct from KNN in their predictive performance trends across different toxicities with a greater balance between sensitivity and specificity scores.

When broken down by descriptor types, the mean F1 scores across all target organ data sets and machine learning algorithms in descending order were: bc 0.70, bct 0.70, bio 0.70, ct 0.68 and chm 0.67. Based on Tukey's HSD post hoc analysis of different descriptor types, differences in performance due to the descriptor types were statistically significant ($p<0.01$) except between bc and bct. The general performance of the different descriptor types had a greater impact on some study types and endpoints compared with others. For example, the type of descriptor had little impact on predicting chronic brain (CHR:Brain) or eye outcomes (CHR:Eye). In contrast, the different descriptors had a greater impact on predicting chronic liver (CHR:Liver) and adrenal outcomes (CHR:Adrenal Gland).

Given the temporal and other design differences in study types we evaluated the trends in target organ toxicity predictions between chronic and subchronic studies. For the same target organ and classification method (i.e., machine learning algorithm, descriptor type and number of descriptors), the F1 scores for subchronic outcomes were generally greater than the equivalent chronic outcomes (i.e., with a difference in mean F1 score of greater than 5%) for the adrenal gland (24%), lungs (24%), thyroid gland (13%), stomach (8%) and spleen (8%). In contrast, chronic outcomes could be predicted more accurately (with a difference in mean F1 score of greater than 5%) for the liver (7%) and thymus (6%). The difference in F1

scores between chronic and subchronic outcomes for the brain, bone marrow, spleen, thymus, and heart were less than 5%.

To systematically examine the impact of different factors on predictive performance, we constructed a fixed effects model to explain the variance in F1 scores based on target organ outcome, study type, descriptor type and machine learning algorithm. The magnitudes of the effects on F1 scores were the highest for target organ toxicity ($\eta^2 = 0.7$) followed by machine learning algorithm ($\eta^2 = 0.2$), descriptor type ($\eta^2 = 0.03$) and study type ($\eta^2 = 0.008$). Based on these results, we decided to further investigate the underlying reasons for the significant effect (η^2) of the machine learning algorithm on the F1 score. In the one-way ANOVA analysis, we found the NB algorithm (mean F1 score=0.61) had the largest contribution among machine learning algorithms on F1 scores. The visualizations in Figure 3 also suggested that the NB algorithm consistently underperformed in comparison to other machine learning algorithms. To reduce the potential bias on the fixed effects modeling, the NB results were excluded from the analysis and the results recalculated. In the updated fixed effects analysis, target organ toxicity still had the greatest effect ($\eta^2 = 0.8$) on F1 scores followed by machine learning algorithm ($\eta^2 = 0.05$) and descriptor type ($\eta^2 = 0.03$). Further analysis of the pairwise interactions between target organ toxicity, machine learning algorithm, and descriptor type showed significant effects on F1 score as follows (in descending order): target organ toxicity and descriptor type ($\eta^2 = 0.02$), target organ toxicity and machine learning algorithm ($\eta^2 = 0.02$), and machine learning algorithm and descriptor type ($\eta^2 = 0.008$). For the minimal data sets, the results suggest that predictive performance was significantly determined by the target organ toxicity outcome being considered followed by the machine learning algorithm, and then the descriptor type.

3. Predictive accuracy using full data sets

The full data sets were used to determine whether the use of additional data changed the estimates of baseline performance or affected the relative impact of different factors on predicting target organ outcomes. Thirty (30) out of the 35 target organ toxicity classes had more than 100 chemicals (i.e., more than 50 positive and 50 negatives). Table 2 shows the performance of the optimal classifiers for each target organ toxicity ranging from a maximum F1 score of 0.85±0.09 (MG:Brain) to a minimum F1 score of 0.67±0.10 (SUB:Kidney). Although the mean F1 score for the full data sets was 0.69 (the same as the minimal data sets), we found some improvements in cases where there were more than 200 chemicals. On average the best F1 scores improved by 17%, sensitivity by 20% and specificity by 11% for 18/35 full data sets in which there were more than 200 chemicals. When aggregated across all descriptor types and machine learning algorithms (excluding NB), we found that the improvement in F1 score, sensitivity and specificity were highly correlated with the number of chemicals (Pearson's ρ= 0.9, 0.7, 0.7, respectively). The top five improvements in F1 scores were observed for subchronic kidney outcomes (29%), chronic kidney outcomes (27%), chronic spleen outcomes (20%), chronic lung outcomes (20%), and chronic adrenal gland outcomes (18%).

When broken down by machine learning algorithm, the mean F1 scores by machine learning algorithms in descending order were: KNN1 0.73, KNN0 0.73, RF0 0.72, CART1 0.71,

CART0 0.71, SVCR0 0.70, SVCL0 0.70 and NB 0.62. The rank order of the machine learning algorithms was the same as that observed for the minimal data sets. However, with more chemicals in the full data set, the more complex classification methods, such as SVC, RF, and CART were used more frequently, and they performed better. Based on Tukey's HSD post hoc analysis, pairwise differences in mean F1 scores between machine learning algorithms were statistically significant ($p<0.01$) except between KNN0 and KNN1. When broken down by descriptor types, the mean F1 scores across all target organ outcomes and machine learning algorithms were, in descending order: bct 0.72, bc 0.72, bio 0.72, ct 0.69 and chm 0.68. The rank order of the relative performance of the descriptor types is also the same as the minimal data sets. Based on Tukey's HSD post hoc analysis, differences in performance due to all descriptor types were statistically significant ($p<0.01$).

To systematically examine the impact of different factors on predictive performance, we constructed a fixed effects model to explain the variance in F1 scores based on target organ toxicity, study type, descriptor type and machine learning algorithm. As in the case of the minimal data sets, we excluded the performance results from the NB machine learning algorithm from the fixed effects model. The effects of the different factors on F1 scores were the greatest for target organ outcome ($\eta^2 = 0.17$), followed by machine learning algorithm ($\eta^2 = 0.26$) and then descriptor type ($\eta^2 = 0.08$). The effect of study type was not significant. Further analysis of the pairwise interactions between target organ toxicity, machine learning algorithm, and descriptor type showed significant effects on F1 score as follows (in descending order): target organ toxicity and descriptor type ($\eta^2 = 0.04$), target organ toxicity and machine learning algorithm ($\eta^2 = 0.03$), and machine learning algorithm and descriptor type ($\eta^2 = 0.01$). In comparison to the fixed effects analysis on the minimal data set, predictive performance was still determined largely by the target organ toxicity outcome being considered. However, the relative effect of target organ outcome was reduced nearly threefold while the effect of machine learning algorithm increased by a similar amount.

4. Relevance of bioactivity and chemotype descriptors for predicting target organ toxicity

We investigated the bioactivity-toxicity and structure-toxicity associations as described in the Methods. Briefly, we constructed the normalized descriptor-toxicity matrix using F1 score $\geq 75^{th}$ percentile, $n_{ds} \leq 25$ and then retained the 50 most frequently used bio descriptors (columns) for predicting chronic target organ toxicity outcomes (rows). The matrix was then hierarchically clustered to putatively group similar organs and bioactivity descriptors, which were then visualized as a heat map (Figure 4). In this heat map, strong positive associations are shown in red. Each row of the heat map shows the relative importance of the bio descriptors for predicting specific chronic target organ toxicity outcomes.

As an illustrative example, we considered a subset of bio descriptors in the heat map (Figure 4) to evaluate their biological relevance to pathological outcomes including: (a) ATG CEBP CIS up and Tox21 PPARg BLA Agonist ratio, and (b) APR HepG2 StressKinase 1h up, APR HepG2 MicrotubuleCSK 72h up and ATG PBREM CIS up, and (c) APR HepG2 CellCycleArrest 72h dn and APR HepG2 CellLoss 24h dn.

In bio descriptor set (a) the “ATG CEBP CIS up” assay measured the increase in activity of the CCAAT/enhancer-binding protein (C/EBP) β via binding to a DNA regulatory region, and “Tox21 PPARg BLA Agonist ratio” measured the change in PPARG activity. The C/EBP family of proteins regulate pro-inflammatory signaling and are involved in tumorigenesis.[42] In addition, the C/EBP transcription factors are highly enriched in the liver where they are involved in the acute phase response.[43] PPARG is primarily known for its role in regulating lipid metabolism. However, it is also expressed in macrophages, where it is involved in inflammation.[44] Inflammation is a physiological response to tissue injury caused by chemical exposure and recognized as a key step in the pathogenesis of chronic diseases in the kidneys[45] and liver.[46] Overall, the two assays in descriptor set (a) could represent markers of inflammation, which is a well-known effect of chemicals on multiple target organs.

Descriptor set (b) contained “APR HepG2 StressKinase 1h up”, “APR HepG2 MicrotubuleCSK 72h up” and “ATG PBREM CIS up”, which measured c-Jun phosphorylation, cytoskeletal stabilization, and activation of the constitutive androstane receptor (CAR), respectively. CAR, which is a member of the nuclear receptor (NR) superfamily, binds to the PBREM and controls the expression of diverse genes including xenobiotic metabolism, liver injury, and hepatocarcinogenesis.[47] Sustained activation of CAR can cause oxidative stress, cellular injury (which can manifest as a cytoskeletal disruption) and regenerative proliferation, which is a key event in hepatocarcinogenesis.[48] The c-Jun protein is part of the AP-1 complex, which is involved in mediating the transcriptional response to oxidative stress.[49] Collectively, descriptor set (b) could represent different markers of cellular stress responses that are induced in the liver, kidneys and multiple other organs.

Finally, bio descriptor set (c) contains “APR HepG2 CellCycleArrest 72h dn” and “APR HepG2 CellLoss 24h dn,” which measured the increase in cell cycle arrest and the cell loss, respectively. Proliferating cells (such as HepG2 cells used in both of these assays) may be unable to progress beyond S-phase in response to stress due to cellular stress and injury.[50] Cell loss can be caused either by the disruption of the cell cycle leading to apoptosis or via other pathways that lead to necrosis. Therefore, assays in descriptor set (c) could be considered as markers of cellular injury and death, which is a phenomenon that is widely observed in multiple target organs.

The bio descriptors found within these three sets are generally overrepresented across a range of chronic toxicity outcomes, but are particularly overrepresented in chronic liver, kidney and spleen toxicities as well as heart, brain, and adrenal responses. The descriptors are consistent with adaptive stress responses, cell injury, cell death, and inflammation playing a major in these pathological outcomes.

We also analyzed structure-toxicity associations using the chemotype (ct) descriptors and visualized the results as a heat map (Figure 5). The strongest associations between ct descriptors (columns) and target organ toxicities (columns) are highlighted in red. We evaluated the ct-tox associations identified by us in terms of already known relationships, or established through reference to structural alerts that are available in typical *in silico* tools

such as the OECD Toolbox[51] and Derek Nexus[52] (within the Nexus 2.1 platform, Lhasa Ltd). Derek Nexus has by far the largest library of organ toxicity alerts. Alerts were available for a number of the toxicity outcomes highlighted here including adrenal gland toxicity, cardiotoxicity, hepatotoxicity and nephrotoxicity. For example, alerts such as "Phenylethyltriazole or analogue" and "2-Thio-benzimidazole, -benzothiazole or -benzoxazole" which were identified for adrenal toxicity from 28 day studies appear to be well aligned with the heteroatom ring chemotypes highlighted in the heat map. There are 97 alerts for heptatoxicity within Derek Nexus, many of them are associated with the pyridine, triazole and 5 membered heteroaromatic chemotypes highlighted in the heat map. Examples include the alert for "Pyrroline or pyrrole ester" (the mechanistic basis of which is discussed in part in the 1988 WHO report[53],http://www.inchem.org/documents/ehc/ehc/ehc080.htm) or "2-Mercaptoimidazole" where one of the lines of evidence underpinning the alert included Mizutani *et al* (2000) who implicated metabolic activation in the liver dysfunction induced by methimazole and related analogues in glutathione (GSH) depleted mice[54]. A systematic assessment of the chemistry to toxicity relationships forms part of our ongoing analysis.

4 DISCUSSION

In this work, we used a number of supervised machine learning to evaluate the utility of *in vitro* bioactivity data from HTS studies and a variety of chemical structure descriptors for predicting 35 target organ toxicity outcomes. Our approach did not utilize any prior knowledge about the relevance of *in vitro assays* to *in vivo* key events, but used an automated and objective approach to systematically explore the impact of five descriptor types, eight classification algorithms, numbers of descriptors and numbers of chemicals on predictive performance for each target organ toxicity. Due to the considerable variation in the numbers of positive and negative chemicals, we created balanced subsets of chemicals for each outcome to reduce any inherent bias in the machine learning modeling. All target organ outcomes (35/35) had minimal balanced data sets with at least 100 chemicals, and many outcomes (30/35) had larger data sets of more than 100 chemicals. We first used the minimal data sets to establish predictive performance baselines for all target organ toxicity outcomes, and then evaluated changes in performance using the full data sets. The mean F1 score across all of these factors was 0.69.

While a predictive performance of 0.69 is far from perfect, it is a substantial improvement over earlier findings based on the smaller set of 309 ToxCast Phase I chemicals.[55] For the minimal data sets, we found KNN classifiers with hybrid descriptors (either bc or bct) produced the most accurate models, which is consistent with previous findings.[22] For some full data sets, SVC, RF, and CART performed better than KNN. Compared to the high sensitivity and low specificity of KNN classifiers, SVC generally offered a greater balance between sensitivity and specificity with a comparable F1 score. This gain in performance of more complex classification algorithms (SVC, RF, and CART) could be attributed to the availability of additional positive and negative chemicals for defining predictive rules about toxicity outcomes. Also, when there were more than 200 chemicals, using the full data sets produced improvements in F1 scores (12%), sensitivity (15%) and specificity (8%). While the performance results for the optimal classification models are promising, we believe that

using the minimal data sets, with the same number of chemicals for each class, provides greater objectivity in comparing the influence of different factors in predicting different types of toxicity outcomes.

We also evaluated the impact of the various factors on performance using fixed effects modeling and found the target organ toxicity outcome was the biggest determinant of predictive performance, followed by the type of descriptor, and then machine learning method. The type of descriptor had a much greater impact on F1 scores in the full data sets ($\eta^2 = 0.152$) as compared to the minimal data sets ($\eta^2 = 0.060$). This suggests, not unexpectedly, that some target organ toxicities may be more difficult to predict than others, [56] but the appropriate choice of descriptors[57] and machine learning algorithm[22] can improve performance. Importantly, bio descriptors in combination with chm or ct descriptors were the most predictive (which is consistent with previous work [57,22]), and chemical structure alone was the least predictive.

Apart from the relative importance of the type of descriptor, we also identified a significant pairwise interaction between toxicity outcome and descriptor type. This suggests that specific endpoints were predicted better with specific types of descriptors. It is possible that the specific types of descriptors contain inherent biases for specific endpoints. For example, ct descriptors may have been constructed using data from a limited number of endpoints while the bio descriptors may be biased towards pathways important in more frequently observed toxicological responses. Greater coverage of biological chemotype space may reduce the level of interaction.

Although the systematic analysis performed in this study identified several important conclusions, the approaches and data sets used in the analysis also have limitations. First, the number of chemicals, number and type of descriptors,[58] and classification algorithm required for accurately predicting a particular toxicity outcome cannot be known *a priori*; they have to be empirically evaluated.[59] When only a handful of chemicals have toxicity data, mining the relationships between descriptors and toxicities is more challenging, and predictive models more likely to be subject to overfitting.[60] Despite a large number of chemicals in our data set, it still represents a very limited sample of environmental chemicals, and it is possible that the results are inflated due to overfitting. Second, although one of our objectives was to compare the impact of different factors (e.g., machine learning algorithms, descriptors types) on predicting diverse target organ toxicities, there are few established statistical approaches for comparing the performance across different classification methods.[61,62] In relatively simple designs when comparing the performance of two different classification methods on a single data set, the mean performance (e.g. F1 score, sensitivity or specificity) and variance are estimated from repetitive random sampling from the same data (i.e. each fold of a cross-validation trial) and can be easily compared. However, multivariate comparisons are much more difficult. Because the samples across cross-validation trials for different classification methods are related, the estimates of performance are not independent and traditional parametric approaches have a high likelihood of detecting differences when there are none.[61] Generally, this is less of an issue when comparing performance scores of different classification methods between unrelated data sets.[62] In addition, in our case, the estimates of performance across data sets may also

be biased because a single chemical can cause multiple target organ toxicities.[20] Due to their high Type I error, we avoided paired t-tests for comparing model performance in favor of traditional analysis of variance techniques (especially since we have a completely balanced data set). Nevertheless, we recognize the challenges in comparing the machine learning results using traditional statistical approaches and hope to address this issue in future work.

Defining a suitable representation for chemicals[63] to predict their toxicity accurately is challenging problem. This is because toxicity, which includes a broad array of abnormal changes in tissue structure or function, can arise via multiple pathways. These pathways can span multiple levels of biological organization (i.e. molecular, cellular, organ), they are highly dynamic and have a complicated relationship with the dose and the duration of chemical exposure [64]. It has been suggested that a finite number of "key" events could determine the outcome of complex pathways.[12] These key events can include receptor activation, gene regulation, or cellular phenotypic changes – all of which may play critical roles in the response of biological systems to chemical exposure. Machine learning provides an unbiased approach to identifying putative biological and chemical descriptors whose higher-order associations map to adverse outcomes.[22] Our preliminary analysis of the bioactivity signatures suggests that machine learning identified a number of plausible key events including adaptive stress responses, cell injury/death, and inflammation. These key events are broadly involved in many pathophysiological processes.[65] Further work is required for identifying key events involved in specific pathways leading to target organ toxicities.

CONCLUSIONS

This work has several important implications and applications for predicting the hazard classifications of new chemicals. First, we have demonstrated that a combination of bioactivity and chemical descriptors can predict a range of target organ toxicity outcomes in chronic, multigenerational, and subchronic guideline studies. These types of predictions are instructive in providing an initial profile of the potential toxicity effects of concern for a chemical which is critical in prioritization as well as in directing analogue identification and evaluation steps in a read-across workflow. Second, we demonstrated that bioactivity descriptors produced more accurate classifiers than the chemical descriptors that were tested. This underscores the importance of continuing to generate HTS data for improving hazard classifications for untested chemicals. Third, the type of descriptor showed significant interaction with the target organ outcome with respect to predictive performance. While certain descriptors may be inherently better at predicting certain toxicological responses, it is likely that the subset of descriptors used in the analysis also contains inherent biases in their construction. For example, most of the 821 *in vitro* assays used in our analysis measured well-known molecular targets in a limited set of cell lines and may over-represent targets involved in a subset of common toxicological responses. Therefore, expanding the biological coverage of the HTS assays, both in terms of the cell types and the molecular targets, may increase the broad predictivity of the machine learning models. Finally, the number of legacy chemicals with published *in vivo* information must be increased (by curation) to reduce the potential for overfitting with limited data. We believe that addressing these issues will

further to improve our ability to predict the target organ toxicities of untested chemicals and reduce our dependence on repeat-dose animal testing experiments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

We are grateful to Sean Ekins and Woodrow Setzer for their helpful feedback on this manuscript.

Funding Sources

This work was supported in part by an appointment to the Research Participation Program at the Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S Department of Energy and EPA.

Abbreviations

5

(Q)SAR	quantitative structure activity relationship
ACEA	ACEA Biosciences, Inc.
ACU	acute toxicity
AOP	adverse outcome pathway
APR	Apredica, Inc.
ATG	Attagene, Inc.
bc	hybrid chemical and bioactivity descriptor
bct	hybrid chemotype and bioactivity descriptor
bio	bioactivity descriptor
C/EBP	CCAAT/enhancer-binding protein
CAR	constitutive androstane receptor
CART	classification and regression trees
chm	chemical description
CHR	chronic toxicity
ct	chemotype descriptor
DEV	developmental toxicity
DNT	developmental neurotoxicity
DSSTox	US EPA Distributed Structure-Searchable Toxicity Database

EPA	Environmental Protection Agency
F1	F1-score
HSD	Tukey's honest significance difference
HTS	high-throughput screening
KNN	k-nearest neighbors
MGR	multigenerational toxicity
MIE	molecular initiating event
NB	naïve Bayes
NRC	National Research Council
NVS	NovaScreen, Inc.
OECD	Organization for Economic Cooperation and Development
OT	Odyssey Thera, Inc.
PBREM	phenobarbital response element
PPARG	peroxisome proliferator-activated receptor gamma
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
REP	reproductive toxicity
RF	random forest
SAC	sub-acute toxicity
SUB	sub-chronic toxicity
SVC	support vector machines classification
tox	toxicity descriptor
ToxCast	US EPA Toxicity Forecaster
ToxRefDB	US EPA Toxicology Reference Database
TSCA	Toxic Substances Control Act
η^2	Eta-squared statistic

7 References

(1). Wagner K, Fach B, and Kolar R (2012) Inconsistencies in data requirements of EU legislation involving tests on animals. ALTEX 29, 302–332. [PubMed: 22847257]

(2). Everts S (2009) Cost Of REACH Underestimated. Chemical and Engineering News 87, 7.

(3). EPA. (2016) TSCA Chemical Substance Inventory, U.S. Environmental Protection Agency.

(4). ECHA. (2016) European Chemicals Agency (ECHA): Pre-registered Substances. Pre-registered substances - ECHA.

(5). Rovida C, and Hartung T (2009) Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals - a report by the transatlantic think tank for toxicology (t(4)). ALTEX 26, 187–208. [PubMed: 19907906]

(6). Committee on Toxicity, T., and Assessment of Environmental Agents, N. R. C. N. (2007) Toxicity Testing in the 21st Century: A Vision and a Strategy. The National Academies Press, Washington, D.C.

(7). Collins FS, Gray GM, and Bucher JR (2008) Toxicology. Transforming environmental health protection. Science 319, 906–907. [PubMed: 18276874]

(8). Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, and Kavlock RJ (2007) The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci 95, 5–12. [PubMed: 16963515]

(9). Kavlock R, and Dix D (2010) Computational toxicology as implemented by the U.S. EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. J Toxicol Environ Health B Crit Rev 13, 197–217. [PubMed: 20574897]

(10). Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, Knudsen T, Martin M, Padilla S, Reif D, Richard A, Rotroff D, Sipes N, and Dix D (2012) Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. Chem Res Toxicol 25, 1287–1302. [PubMed: 22519603]

(11). Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin MT, Wambaugh JF, Knudsen TB, Kancherla J, Mansouri K, Patlewicz G, Williams AJ, Little SB, Crofton KM, and Thomas RS (2016) ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chemical Research in Toxicology.

(12). Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK, Serrrano JA, Tietge JE, and Villeneuve DL (2010) Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ. Toxicol. Chem 29, 730–741. [PubMed: 20821501]

(13). Paul Friedman K, Watt ED, Hornung MW, Hedge JM, Judson RS, Crofton KM, Houck KA, and Simmons SO (2016) Tiered High-Throughput Screening Approach to Identify Thyroperoxidase Inhibitors Within the ToxCast Phase I and II Chemical Libraries. Toxicol Sci 151, 160–180. [PubMed: 26884060]

(14). Browne P, Judson RS, Casey WM, Kleinstreuer NC, and Thomas RS (2015) Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model. Environ Sci Technol 49, 8804–8814. [PubMed: 26066997]

(15). Auerbach S, Filer D, Reif D, Walker V, Holloway AC, Schlezinger J, Srinivasan S, Svoboda D, Judson R, Bucher JR, and Thayer KA (2016) Prioritizing Environmental Chemicals for Obesity and Diabetes Outcomes Research: A Screening Approach Using ToxCast High-Throughput Data. Environ Health Perspect 124, 1141–1154. [PubMed: 26978842]

(16). Knudsen T, Martin M, Chandler K, Kleinstreuer N, Judson R, and Sipes N (2013) Predictive models and computational toxicology. Methods in Molecular Biology (Clifton, N.J.) 947, 343–374.

(17). Kleinstreuer NC, Dix DJ, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Paul KB, Reif DM, Crofton KM, Hamilton K, Hunter R, Shah I, and Judson RS (2013) In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. Toxicol Sci 131, 40–55. [PubMed: 23024176]

(18). Sipes NS, Martin MT, Reif DM, Kleinstreuer NC, Judson RS, Singh AV, Chandler KJ, Dix DJ, Kavlock RJ, and Knudsen TB (2011) Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data. Toxicol Sci 124, 109–127. [PubMed: 21873373]

(19). Martin MT, Knudsen TB, Reif DM, Houck KA, Judson RS, Kavlock RJ, and Dix DJ (2011) Predictive model of rat reproductive toxicity from ToxCast high throughput screening. Biol Reprod 85, 327–339. [PubMed: 21565999]

(20). Martin MT, Judson RS, Reif DM, Kavlock RJ, and Dix DJ (2009) Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. Environ. Health Perspect 117, 392–399. [PubMed: 19337514]

(21). Low Y, Uehara T, Minowa Y, Yamada H, Ohno Y, Urushidani T, Sedykh A, Muratov E, Kuz'min V, Fourches D, Zhu H, Rusyn I, and Tropsha A (2011) Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. Chem Res Toxicol 24, 1251–1262. [PubMed: 21699217]

(22). Liu J, Mansouri K, Judson RS, Martin MT, Hong H, Chen M, Xu X, Thomas RS, and Shah I (2015) Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. Chemical Research in Toxicology 28, 738–751. [PubMed: 25697799]

(23). Fourches D, Muratov E, and Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50, 1189–1204. [PubMed: 20572635]

(24). Williams AJ, and Ekins S (2011) A quality alert and call for improved curation of public chemistry databases. Drug Discov Today 16, 747–750. [PubMed: 21871970]

(25). Williams AJ, Ekins S, and Tkachenko V (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discov Today 17, 685–701. [PubMed: 22426180]

(26). Judson R, Houck K, Martin M, Knudsen T, Thomas RS, Sipes N, Shah I, Wambaugh J, and Crofton K (2014) In vitro and modelling approaches to risk assessment from the U.S. Environmental Protection Agency ToxCast programme. Basic Clin Pharmacol Toxicol 115, 69–76. [PubMed: 24684691]

(27). Martin MT, Judson R, Richard A, Houck KA, and Dix DJ (2007) ToxRefDB: Linking Regulatory Toxicological Information on Environmental Chemicals with High-Throughput Screening and Genomic Data, In International Forum on Computational Toxicology.

(28). Romanov S, Medvedev A, Gambarian M, Poltoratskaya N, Moeser M, Medvedeva L, Gambarian M, Diatchenko L, and Makarov S (2008) Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. Nat Methods 5, 253–260. [PubMed: 18297081]

(29). Knudsen TB, Houck KA, Sipes NS, Singh AV, Judson RS, Martin MT, Weissman A, Kleinstreuer NC, Mortensen HM, Reif DM, Rabinowitz JR, Setzer RW, Richard AM, Dix DJ, and Kavlock RJ (2011) Activity profiles of 309 ToxCast chemicals evaluated across 292 biochemical targets. Toxicology 282, 1–15. [PubMed: 21251949]

(30). Shah I, Setzer RW, Jack J, Houck KA, Judson RS, Knudsen TB, Liu J, Martin MT, Reif DM, Richard AM, Thomas RS, Crofton KM, Dix DJ, and Kavlock RJ (2016) Using ToxCast Data to Reconstruct Dynamic Cell State Trajectories and Estimate Toxicological Points of Departure. Environ Health Perspect 124, 910–919. [PubMed: 26473631]

(31). Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, Trisciuzzi D, Fourches D, Horvath D, Benfenati E, Muratov E, Wedebye EB, Grisoni F, Mangiatordi GF, Incisivo GM, Hong H, Ng HW, Tetko IV, Balabin I, Kancherla J, Shen J, Burton J, Nicklaus M, Cassotti M, Nikolov NG, Nicolotti O, Andersson PL, Zang Q, Politi R, Beger RD, Todeschini R, Huang R, Farag S, Rosenberg SA, Slavov S, Hu X, and Judson RS (2016) CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environ Health Perspect 124, 1023–1033. [PubMed: 26908244]

(32). Rogers D, and Hahn M (2010) Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling 50, 742–754. [PubMed: 20426451]

(33). Landrum G (2015) RDKit.

(34). Yang C, Tarkhov A, Marusczyk J, Bienfait B, Gasteiger J, Kleinoeder T, Magdziarz T, Sacher O, Schwab CH, Schwoebel J, Terfloth L, Arvidson K, Richard A, Worth A, and Rathman J (2015) New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. J Chem Inf Model 55, 510–528. [PubMed: 25647539]

(35). Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, and Steinberg D (2008) Top 10 algorithms in data mining. Knowledge and Information Systems 14, 1–37.

(36). Cortes C, and Vapnik V Support-vector networks. Mach Learn 20, 273–297.

(37). Guyon I, Boser B, and Vapnik V (1993) Automatic Capacity Tuning of Very Large VC-dimension Classifiers, pp 147–155, Morgan Kaufmann.

(38). Breiman L, Friedman J, Stone CJ, and Olshen RA (1984) Classification and Regression Trees. Taylor & Francis.

(39). Tukey JW (1949) Comparing individual means in the analysis of variance. Biometrics, 99–114. [PubMed: 18151955]

(40). Cohen J (1973) Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educational and psychological measurement.

(41). Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9, 90–95.

(42). Balamurugan K, and Sterneck E (2013) The many faces of C/EBPdelta and their relevance for inflammation and cancer. Int J Biol Sci 9, 917–933. [PubMed: 24155666]

(43). Schrem H, Klempnauer J, and Borlak J (2004) Liver-enriched transcription factors in liver function and development. Part II: the C/EBPs and D site-binding protein in cell cycle control, carcinogenesis, circadian gene regulation, liver regeneration, apoptosis, and liver-specific gene regulation. Pharmacol Rev 56, 291–330. [PubMed: 15169930]

(44). Szeles L, Torocsik D, and Nagy L (2007) PPARgamma in immunity and inflammation: cell types and diseases. Biochim Biophys Acta 1771, 1014–1030. [PubMed: 17418635]

(45). Suarez-Alvarez B, Liapis H, and Anders HJ (2016) Links between coagulation, inflammation, regeneration, and fibrosis in kidney pathology. Lab Invest 96, 378–390. [PubMed: 26752746]

(46). Zoller H, and Tilg H (2016) Nonalcoholic fatty liver disease and hepatocellular carcinoma. Metabolism 65, 1151–1160. [PubMed: 26907206]

(47). Kobayashi K, Hashimoto M, Honkakoski P, and Negishi M (2015) Regulation of gene expression by CAR: an update. Arch Toxicol 89, 1045–1055. [PubMed: 25975989]

(48). Kazantseva YA, Pustylnyak YA, and Pustylnyak VO (2016) Role of Nuclear Constitutive Androstane Receptor in Regulation of Hepatocyte Proliferation and Hepatocarcinogenesis. Biochemistry (Mosc) 81, 338–347. [PubMed: 27293091]

(49). Healy S, Khan P, and Davie JR (2013) Immediate early response genes and cell transformation. Pharmacol Ther 137, 64–77. [PubMed: 22983151]

(50). Raza H, John A, and Benedict S (2011) Acetylsalicylic acid-induced oxidative stress, cell cycle arrest, apoptosis and mitochondrial dysfunction in human hepatoma HepG2 cells. Eur J Pharmacol 668, 15–24. [PubMed: 21722632]

(51). Dimitrov SD, Diderich R, Sobanski T, Pavlov TS, Chankov GV, Chapkanov AS, Karakolev YH, Temelkov SG, Vasilev RA, Gerova KD, Kuseva CD, Todorova ND, Mehmed AM, Rasenberg M, and Mekenyan OG (2016). QSAR Toolbox - workflow and major functionalities. SAR QSAR Environ Res. 19, 1–17.

(52). Bhhatari B, Wilson DM, Parks AK, Carnery EW, and Spencer PJ (2016) Evaluation of TOPKAT, Toxtree, and Derek Nexus in Silicol Models for Ocular Irritation and Development of a Knowledge-Based Framework to Improve the Prediction of Severe Irritation. Chem Res Toxicol 29, 810–822. [PubMed: 27018716]

(53). (WHO), W. H. O. (1988) Pyrrolizidine alkaloids, In International Programme on Chemical Safety, Geneva.

(54). Mizutani T, Yoshida K, Murakami M, Shirai M, and Kawazoe S (200) Evidence for the involvement of N-methylthiourea, a ring cleavage metabolite in the hepatotoxicity of methimazole in glutathione-depleted mice: structure-toxicity and metabolic studies. Chem Res Toxicol 13, 170–176. [PubMed: 10725113]

(55). Thomas RS, Black MB, Li L, Healy E, Chu TM, Bao W, Andersen ME, and Wolfinger RD (2012) A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. Toxicol Sci 128, 398–417. [PubMed: 22543276]

(56). Ekins S (2014) Progress in computational toxicology. J Pharmacol Toxicol Methods 69, 115–140. [PubMed: 24361690]

(57). Low Y, Sedykh A, Fourches D, Golbraikh A, Whelan M, Rusyn I, and Tropsha A (2013) Integrative chemical-biological read-across approach for chemical hazard classification. Chemical Research in Toxicology 26, 1199–1208. [PubMed: 23848138]

(58). Blum AL, and Langley P (1997) Selection of relevant features and examples in machine learning. Artificial intelligence 97, 245–271.

(59). Weiss SM, and Kapouleas I (1990) An empirical comparison of pattern recognition, neural nets and machine learning classification methods. Readings in machine learning, 177–183.

(60). Dietterich T (1995) Overfitting and undercomputing in machine learning. ACM computing surveys (CSUR) 27, 326–327.

(61). Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation 10, 1895–1923. [PubMed: 9744903]

(62). Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research 7, 1–30.

(63). Sahoo S, Adhikari C, Kuanar M, and Mishra BK (2016) A Short Review of the Generation of Molecular Descriptors and Their Applications in Quantitative Structure Property/Activity Relationships. Curr Comput Aided Drug Des.

(64). Edwards SW, and Preston RJ (2008) Systems biology and mode of action based risk assessment. Toxicol Sci 106, 312–318. [PubMed: 18791183]

(65). Becker RA, Ankley GT, Edwards SW, Kennedy SW, Linkov I, Meek B, Sachana M, Segner H, Van Der Burg B, Villeneuve DL, Watanabe H, and Barton-Maclaren TS (2015) Increasing Scientific Confidence in Adverse Outcome Pathways: Application of Tailored Bradford-Hill Considerations for Evaluating Weight of Evidence. Regul Toxicol Pharmacol 72, 514–537. [PubMed: 25863193]

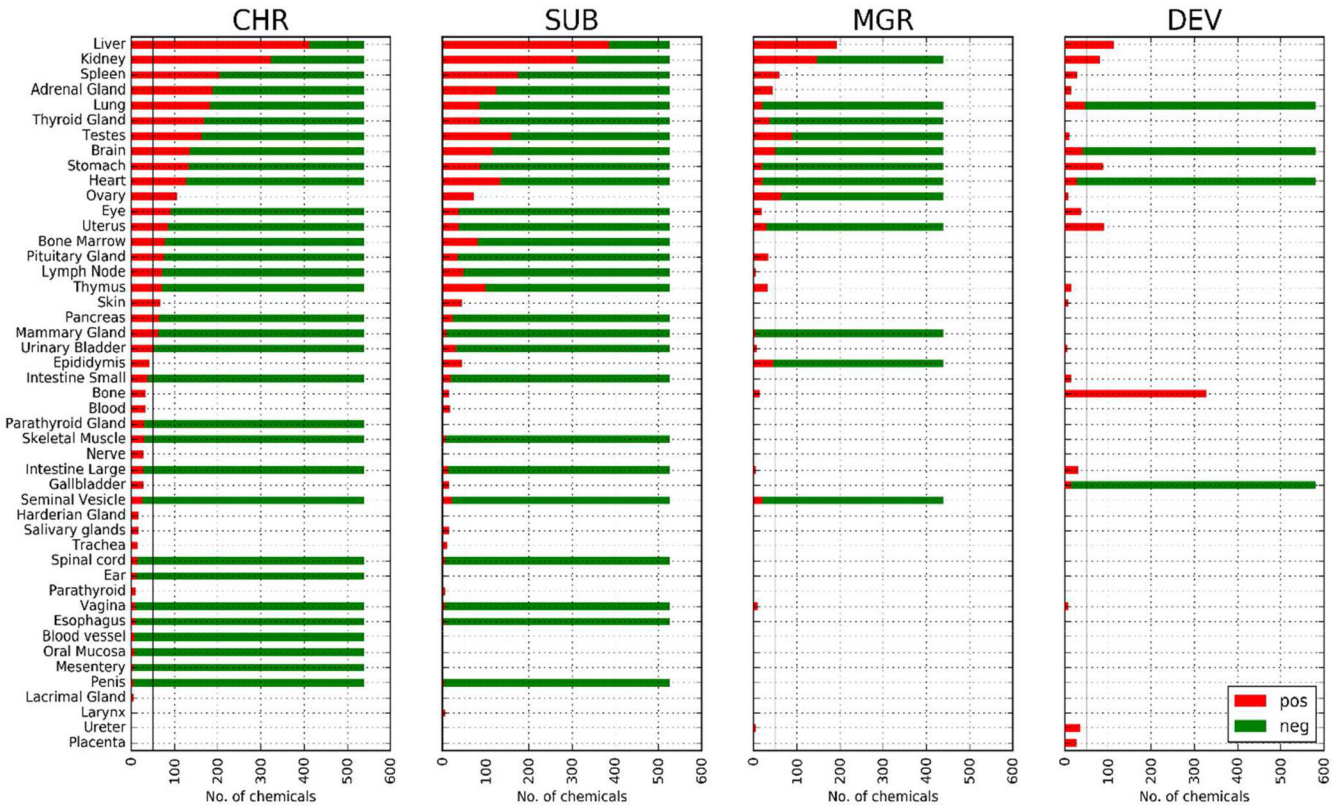


Figure 1.

Distribution of positive and negative chemicals across the *in vivo* guideline toxicity testing studies and target organs. From left to right these bar graphs show the number of positive (pos, red) and negative (neg, green) chemicals for chronic (CHR), subchronic (SUB), multigenerational (MGR) and developmental (DEV) studies. The target organs are labeled on the ordinate and the number of chemicals on the abscissa. The negative chemicals are missing for guideline studies where the evaluation of the specific target organ effect was not compulsory.

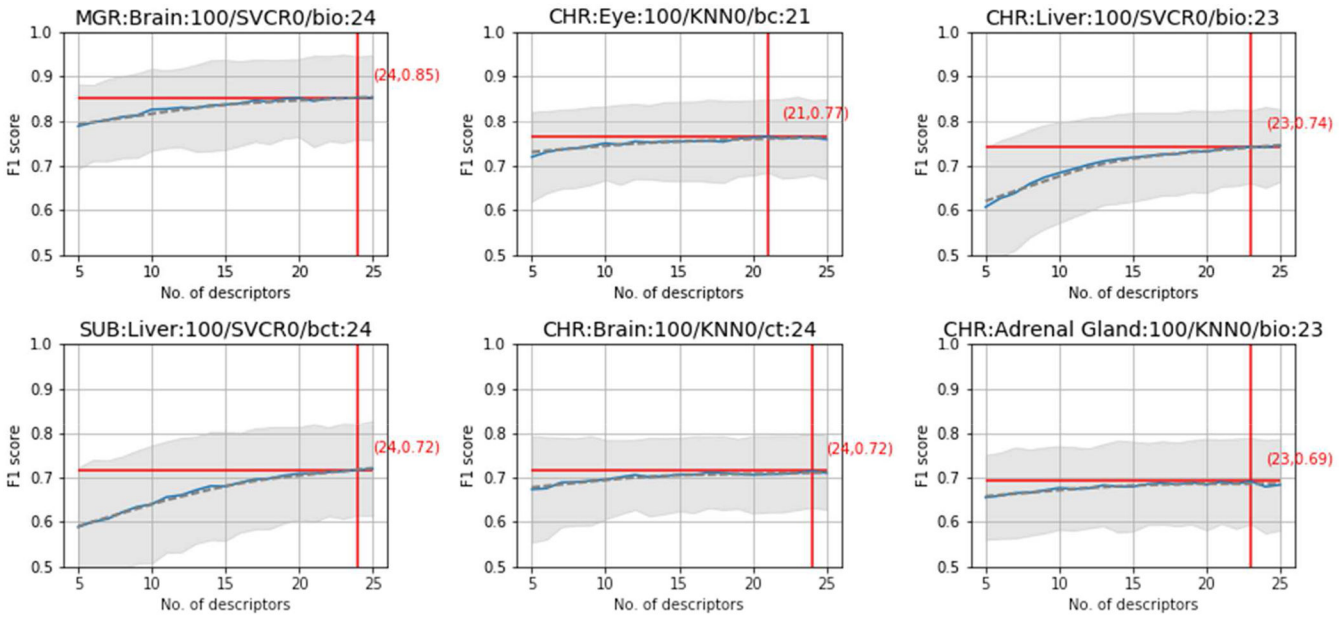


Figure 2.

Relationship between F1 score and number of descriptors for the best performing classification models and illustrative examples of minimal datasets. In each graph, the effect and descriptor type are given in the title (denoted as study:target-organ), the mean F1 score, and the standard deviation is shown in blue and gray, respectively. The number of descriptors and F1 score for the best classifier are signified on the ordinate and abscissa, respectively, by vertical and horizontal red lines. Each graph shows the cross-validation F1 score (ordinate) and number of descriptors (abscissa) for predicting toxicities (shown in the title and denoted as study:target-organ) using classification methods (shown in title)

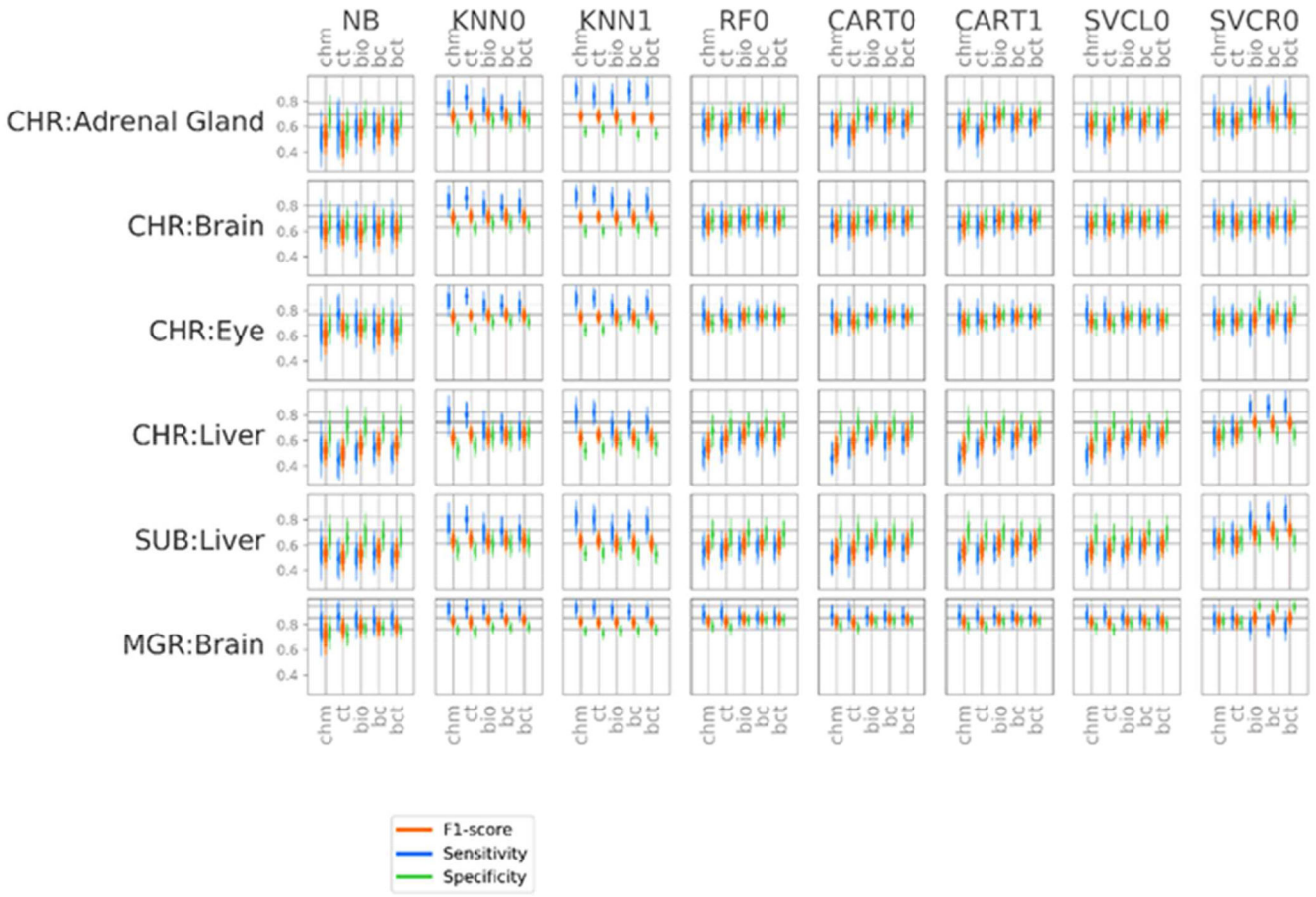


Figure 3.

Summary of performance for target organ outcomes for select minimum datasets by classification algorithms and descriptors. The visualization shows the predictive performance for illustrative examples of target organ outcomes in rows (denoted as, study:target organ) using eight machine learning algorithms (columns): naïve Bayes (NB), k-nearest neighbor classification (KNN0 and KNN1) classification and regression trees (CART0 and CART1) and support vector classifiers (SVCL0 and SVCR0). The predictive performance is compared across five different descriptors including: chemical (chm), chemotype (ct), *in vitro* bioactivity (bio), a combination of *in vitro* bioactivity and chemical (bc), and a combination of *in vitro* bioactivity and chemotype (ct). The performance of a classification method for predicting an outcomeusing a descriptor type was measured using specificity (green), F1 score (red) and sensitivity (blue), which are visualized as vertical glyphs. The center, top, and bottom of the glyphs correspond to the mean ±1 SD. In all, the performance results for 40 classification methods (8 machine learning algorithms and five descriptor types) are visualized for each target organ toxicity. The grey horizontal bars on each graph signify the best mean F1 score ±1 SD (across all 5-fold cross-validation trials). The best performing classification model and descriptor set for each target organ outcome are denoted with a vertical red line.

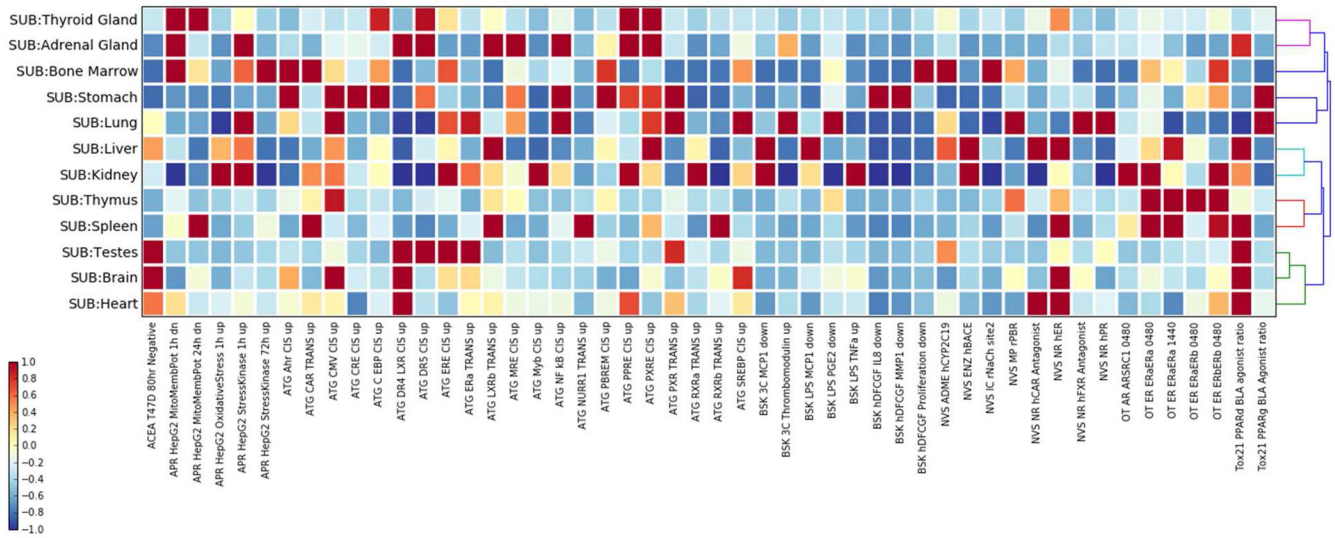


Figure 4.

Summary of frequently used bioactivity descriptors in chronic target organ toxicity prediction models. The visualization shows a heat map in which the rows correspond to chronic target organ toxicities, columns correspond to the fifty most frequently used bioactivity descriptors, and values represent row standardized frequencies of occurrence of descriptors (column) in predictive models of target organ toxicities (row). The colors signify the row standardized frequencies for the bioactivity descriptors where positive values are red, negative values are blue and the level of saturation is directly related to magnitude. The row dendrogram show the cosine similarity between the frequency of bioactivity descriptors and target organ toxicity outcomes, respectively, by average linkage clustering.

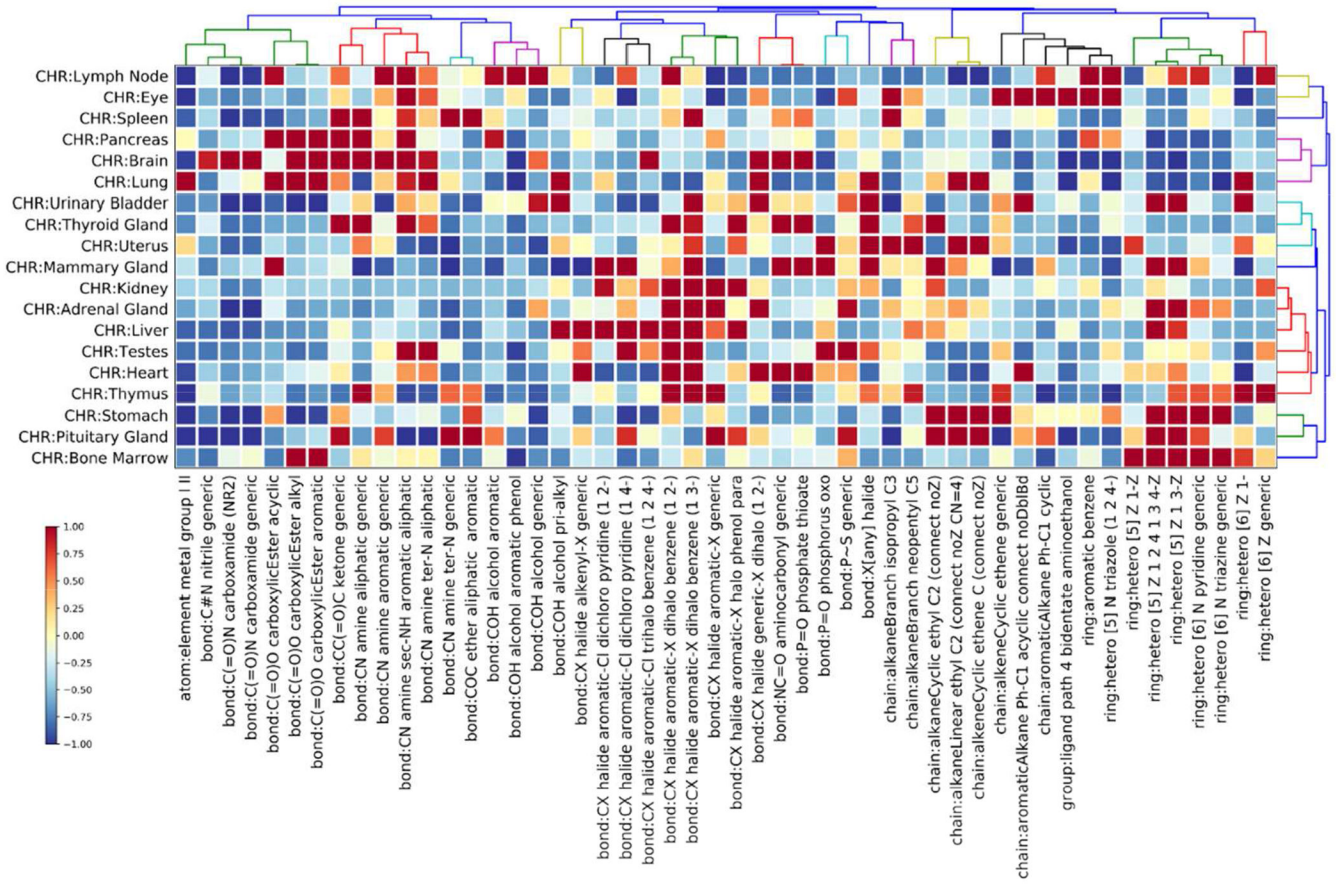


Figure 5.

Summary of frequently used chemotype descriptors in chronic target organ toxicity prediction models. The visualization shows a heat map in which the rows correspond to chronic target organ toxicities, columns correspond to the fifty most frequently used chemotype descriptors, and values represent row standardized frequencies of occurrence of descriptors (column) in predictive models of target organ toxicities (row). The colors signify the row standardized frequencies for the chemotype descriptors where positive values are red, negative values are blue and the level of saturation is directly related to magnitude. The row dendrogram shows the cosine similarity between the frequency of chemotype descriptors and target organ toxicity outcomes, respectively, by average linkage clustering.

Table 1.

Performance baseline of optimal classifiers for the minimal data sets (50 positive and 50 negative chemicals). From left to right the columns show the classifier identified (Classifier id), the organ toxicity (denoted as study:target organ), the machine learning algorithm (Algorithm), the descriptor type (dt), the number of descriptors used (n_{ds}), the F1 score ± 1 SD, sensitivity± 1 SD and specificity± 1 SD.

Classifier id	Organ Toxicity	Algorithm	dt	n_ds	F1 Score	Sensitivity	Specificity
MGR:Brain:100/SVCR0/bio:24	MGR:Brain	SVCR0	bio	24	0.85±0.09	0.79±0.13	0.95±0.06
CHR:Urinary Bladder:100/SVCR0/bct:24	CHR:Urinary Bladder	SVCR0	bct	24	0.85±0.09	0.80±0.13	0.93±0.07
CHR:Mammary Gland:100/SVCR0/bc:24	CHR:Mammary Gland	SVCR0	bc	24	0.83±0.10	0.77±0.14	0.93±0.07
CHR:Pancreas:100/KNN0/bc:22	CHR:Pancreas	KNN0	bc	22	0.81±0.07	0.89±0.10	0.75±0.07
SUB:Bone Marrow:100/KNN0/bct:13	SUB:Bone Marrow	KNN0	bct	13	0.81±0.07	0.90±0.10	0.74±0.07
CHR:Lymph Node:100/KNN0/bct:24	CHR:Lymph Node	KNN0	bct	24	0.81±0.07	0.89±0.10	0.74±0.07
MGR:Ovary:100/KNN0/bc:24	MGR:Ovary	KNN0	bc	24	0.81±0.08	0.87±0.11	0.75±0.07
CHR:Thymus:100/KNN0/bc:22	CHR:Thymus	KNN0	bc	22	0.80±0.08	0.87±0.11	0.75±0.08
CHR:Bone Marrow:100/KNN0/bct:24	CHR:Bone Marrow	KNN0	bct	24	0.79±0.07	0.89±0.10	0.72±0.07
CHR:Uterus:100/KNN0/bct:23	CHR:Uterus	KNN0	bct	23	0.79±0.08	0.86±0.12	0.73±0.08
SUB:Lung:100/KNN0/ct:23	SUB:Lung	KNN0	ct	23	0.78±0.07	0.91±0.10	0.69±0.07
MGR:Testes:100/KNN0/bct:24	MGR:Testes	KNN0	bct	24	0.78±0.08	0.86±0.12	0.72±0.08
SUB:Stomach:100/KNN0/ct:22	SUB:Stomach	KNN0	ct	22	0.78±0.07	0.90±0.11	0.69±0.07
CHR:Pituitary Gland:100/KNN0/bio:24	CHR:Pituitary Gland	KNN0	bio	24	0.77±0.08	0.86±0.12	0.71±0.08
SUB:Thyroid Gland:100/KNN0/bio:22	SUB:Thyroid Gland	KNN0	bio	22	0.77±0.08	0.86±0.11	0.71±0.09
CHR:Eye:100/KNN0/bc:21	CHR:Eye	KNN0	bc	21	0.77±0.08	0.84±0.12	0.72±0.08
SUB:Thymus:100/KNN0/bio:22	SUB:Thymus	KNN0	bio	22	0.77±0.09	0.86±0.13	0.70±0.08
SUB:Adrenal Gland:100/KNN0/bc:22	SUB:Adrenal Gland	KNN0	bc	22	0.76±0.09	0.82±0.13	0.71±0.09
CHR:Heart:100/KNN0/bct:21	CHR:Heart	KNN0	bct	21	0.75±0.09	0.81±0.13	0.72±0.09
SUB:Brain:100/KNN0/bio:24	SUB:Brain	KNN0	bio	24	0.75±0.09	0.83±0.13	0.69±0.08
SUB:Heart:100/KNN0/bc:22	SUB:Heart	KNN0	bc	22	0.74±0.09	0.82±0.13	0.69±0.08
CHR:Liver:100/SVCR0/bio:23	CHR:Liver	SVCR0	bio	23	0.74±0.08	0.86±0.11	0.66±0.09
CHR:Stomach:100/KNN0/bct:23	CHR:Stomach	KNN0	bct	23	0.74±0.09	0.81±0.13	0.69±0.09
SUB:Testes:100/KNN0/bc:23	SUB:Testes	KNN0	bc	23	0.74±0.09	0.83±0.13	0.67±0.08
MGR:Kidney:100/KNN0/ct:22	MGR:Kidney	KNN0	ct	22	0.73±0.08	0.88±0.12	0.63±0.08
SUB:Liver:100/SVCR0/bct:24	SUB:Liver	SVCR0	bct	24	0.72±0.10	0.85±0.15	0.64±0.10
CHR:Brain:100/KNN0/ct:24	CHR:Brain	KNN0	ct	24	0.72±0.08	0.86±0.13	0.62±0.08
CHR:Thyroid Gland:100/KNN0/ct:23	CHR:Thyroid Gland	KNN0	ct	23	0.71±0.08	0.86±0.13	0.62±0.08
SUB:Spleen:100/KNN0/bio:24	SUB:Spleen	KNN0	bio	24	0.70±0.09	0.79±0.14	0.65±0.09
CHR:Lung:100/KNN0/bct:19	CHR:Lung	KNN0	bct	19	0.70±0.09	0.79±0.14	0.64±0.09
CHR:Testes:100/KNN0/ct:22	CHR:Testes	KNN0	ct	22	0.70±0.08	0.85±0.13	0.60±0.08
CHR:Adrenal Gland:100/KNN0/bio:23	CHR:Adrenal Gland	KNN0	bio	23	0.69±0.10	0.77±0.15	0.65±0.10
CHR:Kidney:100/SVCR0/bio:24	CHR:Kidney	SVCR0	bio	24	0.69±0.12	0.83±0.18	0.62±0.11
CHR:Spleen:100/KNN0/ct:23	CHR:Spleen	KNN0	ct	23	0.68±0.08	0.85±0.12	0.58±0.08

Classifier id	Organ Toxicity	Algorithm	dt	n_ds	F1 Score	Sensitivity	Specificity
SUB:Kidney:100/KNN0/bio:20	SUB:Kidney	KNN0	bio	20	0.67±0.10	0.77±0.14	0.61±0.10

Machine learning algorithms: Naïve Bayes (NB), support vector classifiers with radial basis function kernel (SVCR0), k-nearest neighbors (KNN0/k=3 and KNN1/k=5), and random forest (RF0).

Table 2.

Performance baseline of optimal classifiers for the full data sets. From left to right the columns show the classifier identified (Classifier id), the organ toxicity (denoted as study:target organ), the machine learning algorithm (Algorithm), the descriptor type (dt), the number of chemicals (n_{chm}), the number of descriptors used (n_{ds}), the F1 score ± 1 SD, sensitivity± 1 SD and specificity± 1 SD.

Classifier id	Organ Toxicity	Algorithm	dt	n_ds	F1 Score	Sensitivity	Specificity
SUB:Bone Marrow:164/SVCR0/bio:23	SUB:Bone Marrow	SVCR0	bio	23	0.88±0.06	0.87±0.09	0.90±0.05
CHR:Mammary Gland:128/SVCR0/bc:23	CHR:Mammary Gland	SVCR0	bc	23	0.88±0.07	0.85±0.10	0.93±0.06
CHR:Urinary Bladder:106/SVCR0/bc:24	CHR:Urinary Bladder	SVCR0	bc	24	0.87±0.08	0.82±0.12	0.93±0.06
CHR:Pancreas:130/SVCR0/bct:23	CHR:Pancreas	SVCR0	bct	23	0.87±0.07	0.82±0.11	0.93±0.06
SUB:Lung:170/SVCR0/bc:23	SUB:Lung	SVCR0	bc	23	0.87±0.06	0.85±0.09	0.89±0.06
MGR:Testes:182/SVCR0/bct:23	MGR:Testes	SVCR0	bct	23	0.86±0.06	0.83±0.09	0.91±0.05
CHR:Heart:250/SVCR0/bio:24	CHR:Heart	SVCR0	bio	24	0.86±0.06	0.80±0.09	0.96±0.05
CHR:Uterus:170/SVCR0/bct:23	CHR:Uterus	SVCR0	bct	23	0.86±0.07	0.80±0.11	0.94±0.06
MGR:Brain:102/SVCR0/bio:23	MGR:Brain	SVCR0	bio	23	0.86±0.09	0.80±0.13	0.94±0.06
SUB:Brain:230/SVCR0/bct:24	SUB:Brain	SVCR0	bct	24	0.86±0.06	0.81±0.09	0.92±0.05
CHR:Bone Marrow:156/SVCR0/bct:24	CHR:Bone Marrow	SVCR0	bct	24	0.85±0.07	0.81±0.11	0.91±0.06
SUB:Thymus:200/SVCR0/bio:24	SUB:Thymus	SVCR0	bio	24	0.85±0.06	0.82±0.09	0.90±0.06
MGR:Ovary:130/SVCR0/bct:24	MGR:Ovary	SVCR0	bct	24	0.85±0.08	0.80±0.12	0.92±0.07
SUB:Thyroid Gland:176/RF0/bct:24	SUB:Thyroid Gland	RF0	bct	24	0.85±0.06	0.86±0.09	0.85±0.06
MGR:Kidney:290/RF0/bct:24	MGR:Kidney	RF0	bct	24	0.85±0.05	0.87±0.07	0.83±0.05
CHR:Lymph Node:140/SVCR0/bc:24	CHR:Lymph Node	SVCR0	bc	24	0.85±0.08	0.79±0.11	0.94±0.06
SUB:Testes:318/RF0/bio:24	SUB:Testes	RF0	bio	24	0.85±0.04	0.89±0.06	0.81±0.05
SUB:Adrenal Gland:250/RF0/bct:24	SUB:Adrenal Gland	RF0	bct	24	0.84±0.05	0.85±0.08	0.84±0.05
CHR:Eye:186/SVCR0/bio:24	CHR:Eye	SVCR0	bio	24	0.84±0.07	0.77±0.11	0.94±0.06
SUB:Heart:268/SVCR0/bio:24	SUB:Heart	SVCR0	bio	24	0.84±0.06	0.82±0.09	0.87±0.06
SUB:Stomach:176/KNN0/bc:23	SUB:Stomach	KNN0	bc	23	0.84±0.05	0.92±0.08	0.78±0.06
CHR:Stomach:268/SVCR0/bc:24	CHR:Stomach	SVCR0	bc	24	0.84±0.06	0.82±0.10	0.86±0.06
CHR:Thymus:140/SVCR0/bct:24	CHR:Thymus	SVCR0	bct	24	0.84±0.08	0.77±0.11	0.93±0.06
SUB:Spleen:352/RF0/bio:24	SUB:Spleen	RF0	bio	24	0.84±0.05	0.85±0.07	0.83±0.05
CHR:Testes:320/KNN0/bct:24	CHR:Testes	KNN0	bct	24	0.83±0.04	0.92±0.06	0.76±0.05
CHR:Brain:260/SVCR0/bct:24	CHR:Brain	SVCR0	bct	24	0.83±0.06	0.78±0.10	0.90±0.06
CHR:Thyroid Gland:320/RF0/bio:24	CHR:Thyroid Gland	RF0	bio	24	0.83±0.05	0.83±0.08	0.84±0.05
CHR:Pituitary Gland:152/SVCR0/bio:23	CHR:Pituitary Gland	SVCR0	bio	23	0.83±0.08	0.79±0.12	0.90±0.08
CHR:Spleen:410/RF0/bio:24	CHR:Spleen	RF0	bio	24	0.83±0.05	0.84±0.07	0.83±0.05
CHR:Lung:366/KNN0/bct:24	CHR:Lung	KNN0	bct	24	0.83±0.04	0.92±0.07	0.76±0.05
CHR:Liver:240/SVCR0/bio:24	CHR:Liver	SVCR0	bio	24	0.83±0.06	0.88±0.07	0.79±0.08
CHR:Adrenal Gland:376/RF0/bct:24	CHR:Adrenal Gland	RF0	bct	24	0.83±0.05	0.81±0.08	0.84±0.05
CHR:Kidney:350/SVCR0/bct:24	CHR:Kidney	SVCR0	bct	24	0.82±0.05	0.87±0.07	0.78±0.06
SUB:Liver:230/SVCR0/bct:24	SUB:Liver	SVCR0	bct	24	0.79±0.07	0.83±0.09	0.77±0.08

Classifier id	Organ Toxicity	Algorithm	dt	n_ds	F1 Score	Sensitivity	Specificity
SUB:Kidney:428/KNN0/bio:24	SUB:Kidney	KNN0	bio	24	0.78±0.05	0.87±0.08	0.71±0.07

Machine learning algorithms: Naïve Bayes (NB), support vector classifiers with radial basis function kernel (SVCR0), k-nearest neighbors (KNN0/k=3 and KNN1/k=5), random forest (RF0) and classification and regression trees (CART0/max-depth=5, CART1/max-depth=automatic).