HHS Public Access

Author manuscript

Comput Toxicol. Author manuscript; available in PMC 2021 May 18.

Published in final edited form as:

Comput Toxicol. 2021 May ; 18: . doi:10.1016/j.comtox.2021.100166.

Predictive modeling of biological responses in the rat liver using *in vitro* Tox21 bioactivity: Benefits from high-throughput toxicokinetics

Caroline Ring[a], **Nisha S. Sipes**[b], **Jui-Hua Hsieh**[c], **Celeste Carberry**[d,e], **Lauren E. Koval**[d,e], **William D. Klaren**[f], **Mark A. Harris**[g], **Scott S. Auerbach**[b], **Julia E. Rager**[d,e,h,*]

[a]ToxStrategies, Inc., Austin, TX 78751, United States

[b]Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, United States

[c]Kelly Government Solutions, Durham, NC 27709, United States

[d]Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States

[e]The Institute for Environmental Health Solutions, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States

[f]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77840, United States

[g]ToxStrategies, Inc., Houston, TX 77494, United States

[h]Curriculum in Toxicology and Environmental Medicine, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States

Abstract

*Corresponding author at: The University of North Carolina at Chapel Hill, 135 Dauer Drive, Chapel Hill, NC 27599, United States. jrager@unc.edu (J.E. Rager).

CRediT authorship contribution statement

Caroline Ring: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Nisha S. Sipes:** Formal analysis, Methodology, Resources, Software, Validation, Writing - review & editing. **Jui-Hua Hsieh:** Data curation, Methodology, Resources, Software, Writing - review & editing. **Celeste Carberry:** Data curation, Visualization, Writing - original draft. **Lauren E. Koval:** Data curation, Software, Validation. **William D. Klaren:** Data curation, Writing - original draft. **Mark A. Harris:** Funding acquisition, Project administration, Writing - review & editing. **Scott S. Auerbach:** Conceptualization, Methodology, Resources, Software, Validation, Writing - review & editing. **Julia E. Rager:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A Supplementary material

All supplemental material are available through UNC Dataverse, located at: Ring C, Sipes NS, Hsieh JH, Carberry C, Koval LE, Klaren WD, Harris MA, Auerbach SS, Rager JE, Dataset for Predictive Modeling of Biological Responses in the Rat Liver using In Vitro Tox21 Bioactivity: Benefits from High-Throughput Toxicokinetics, UNC Dataverse, Ragerlab-Dataverse. 2020. Available at: https://doi.org/10.15139/S3/WCLFWZ.

Computational methods are needed to more efficiently leverage data from *in vitro* cell-based models to predict what occurs within whole body systems after chemical insults. This study set out to test the hypothesis that *in vitro* high-throughput screening (HTS) data can more effectively predict *in vivo* biological responses when chemical disposition and toxicokinetic (TK) modeling are employed. *In vitro* HTS data from the Tox21 consortium were analyzed in concert with chemical disposition modeling to derive nominal, aqueous, and intracellular estimates of concentrations eliciting 50% maximal activity. *In vivo* biological responses were captured using rat liver transcriptomic data from the DrugMatrix and TG-Gates databases and evaluated for pathway enrichment. *In vivo* dosing data were translated to equivalent body concentrations using HTTK modeling. Random forest models were then trained and tested to predict *in vivo* pathway-level activity across 221 chemicals using *in vitro* bioactivity data and physicochemical properties as predictor variables, incorporating methods to address imbalanced training data resulting from high instances of inactivity. Model performance was quantified using the area under the receiver operator characteristic curve (AUC-ROC) and compared across pathways for different combinations of predictor variables. All models that included toxicokinetics were found to outperform those that excluded toxicokinetics. Biological interpretation of the model features revealed that rather than a direct mapping of *in vitro* assays to *in vivo* pathways, unexpected combinations of multiple *in vitro* assays predicted *in vivo* pathway-level activities. To demonstrate the utility of these findings, the highest-performing model was leveraged to make new predictions of *in vivo* biological responses across all biological pathways for remaining chemicals tested in Tox21 with adequate data coverage (n = 6617). These results demonstrate that, when chemical disposition and toxicokinetics are carefully considered, *in vitro* HT screening data can be used to effectively predict *in vivo* biological responses to chemicals.

Keywords

Predictive modeling; Machine learning; ToxCast/Tox21; Transcriptomics; Biological pathways; Toxicokinetics

1. Introduction

Computational approaches are currently being developed to more effectively interpret and translate biological activity measured *in vitro* using cell-based models to predict *in vivo* biological responses and ultimately inform health outcomes in humans. These approaches are needed throughout many areas of science, including molecular biology, medicine, toxicology, and environmental science. As an example, there is a need to accelerate the identification and testing of chemicals for efficacy and safety during the development of pharmaceuticals [1,2].

Within the chemical industry and regulatory agencies, there is a parallel push to decrease the time required to identify deleterious chemicals and reduce reliance upon animals in chemical safety testing [3–6]. Key to this movement's success is the effective incorporation of *in vitro* testing, in which there is a current drive to advance computational methods that can be used to more confidently translate *in vitro* findings into the context of *in vivo* biology.

The field of computational biology and toxicology is rapidly expanding to address this issue by developing methods that use *in vitro* data, in combination with *in silico* approaches, to predict *in vivo* toxicity responses [7–12]. Computational methods aimed at predicting *in vivo* toxicity outcomes continue to integrate recent advancements in machine learning algorithms alongside increasing amounts of information that can be used to build and test predictive models [7]. In terms of *in vitro* data, predictive models commonly incorporate data from high-throughput screening (HTS) efforts to maximize chemical coverage, with one of the most recognized HTS examples being the ToxCast/Tox21 consortium [13]. Other data that have been used to aid in overall predictions include chemical structure descriptors, physiochemical properties, and assay descriptors, among others [7,14,15]. Variables that are often used as the outcomes include toxicological changes observed at the pathology and/or disease-level [15–17]. These outcomes can also be gathered from databases to cover wider chemical domains, such as the ToxRefDB [18]. With these data, predictive models are trained and tested with the goal of deriving computational models that can be used to provide a prediction of whether a chemical will elicit a certain toxicological outcome based on *in vitro* findings.

Predictive modeling efforts based on HTS data from the ToxCast/Tox21 program have shown variable results, to date. Previous studies have integrated ToxCast/Tox21 bioactivity profiles into predictive models by using general *in vitro* activity calls (i.e., active vs. inactive) and/or concentrations identified to elicit *in vitro* activity (e.g., concentrations eliciting 50% maximal activity) as the primary predictor variables [14–17,19,20]. Though serving as important starting points, these studies have shown mixed results, with many demonstrating limited success [15,16,19,20]. In an effort to improve model predictivity, we implement a novel approach to build models that predict mechanisms of *in vivo* toxicity, rather than higher level outcomes at the pathology/disease-level. Our approach further benefits from the testing of more precise chemical disposition and toxicokinetic parameters to more accurately translate *in vitro* bioactivity concentrations into the context of *in vivo* biology.

Predictive toxicology efforts have only recently started incorporating improved estimates of chemical disposition and toxicokinetics. A notable study recently demonstrated that applying physiologically based toxicokinetic modeling improved the ability to associate *in vitro* ToxCast bioactivity to rat toxicity endpoints gathered from ToxRefDB and animal testing data aggregated in the U.S. EPA's computational toxicology dashboard [21]. Another recent study implemented *in vitro* to *in vivo* extrapolation methods incorporating toxicokinetics to relate points of departure obtained using *in vitro* bioactivity against those derived using traditional *in vivo* hazard information [22]. We also recently demonstrated that concordance between *in vitro* and *in vivo* biological responses significantly improves after toxicokinetics are accounted for [23]. The current study builds on these data and serves as the first to leverage recent advancements in chemical disposition and toxicokinetic modeling to better extract biologically meaningful *in vitro* chemical-response activity profiles and use these to predict *in vivo* biological responses, focusing on pathway alterations within the rat liver.

2. Materials and methods

2.1. Study overview

This study set out to evaluate whether including chemical disposition and toxicokinetic modeling improves the ability to use *in vitro* HTS data to predict *in vivo* biological responses to chemical exposures (Fig. 1). Machine-learning models were trained to infer relationships between *in vitro* activity concentrations and whether *in vivo* pathway-level responses were elicited at a particular dose. These relationships may be modified by considering chemical disposition *in vitro*, which links the nominal chemical concentration placed into an *in vitro* well to the concentrations that occur as the chemical partitions into various parts of the assay system (Fig. 1A). These relationships may also be modified by considering toxicokinetics, which links external *in vivo* doses to internal body concentrations by describing the body's absorption, distribution, metabolism, and excretion of a chemical (Fig. 1B).

Depending on the mechanism of *in vitro* bioactivity, various chemical-disposition metrics may be relevant (Fig. 1A). *In vitro* activity may be associated with the intracellular concentration in the assay system, or it may be associated with the concentration in the aqueous phase of the assay system. Similarly, depending on the mechanism of *in vivo* activity, various TK-predicted metrics of internal body concentration may be relevant. First, *in vivo* activity may be associated with concentration in a particular target tissue (e.g. liver), or concentration in circulating plasma (Fig. 1B, **left**). Second, *in vivo* activity may be influenced by time-varying concentrations in different ways: Activity may occur when a tissue concentration reaches a particular maximum level (even if that level is not sustained for long), or it may occur only when a tissue concentration achieves a certain average level over time (Fig. 1B, **right**).

Predictive performance was therefore compared for machine-learning models including eight different combinations of TK metrics and chemical-disposition metrics as predictor variables: maximum and mean liver concentration, each combined with nominal and intracellular AC_{50} values; maximum and mean plasma concentration, each combined with nominal and aqueous-phase AC_{50} values. The performance of these models was also compared to machine-learning models that excluded TK metrics (only *in vivo* external dose combined with nominal, intracellular, or aqueous-phase AC_{50}) and models that included neither TK nor chemical-disposition metrics. In order to directly evaluate whether TK metrics of internal dose improved model performance above *in vivo* external dose alone, *in vivo* external dose was included in all models. Table 1 details the twelve machine-learning models that were built for each pathway and compared. As a result of these model comparisons, this study evaluated whether TK improves model performance at predicting *in vivo* pathway-level activity compared to external dose alone, and if so, which TK-predicted metric yields the best performance.

The following steps were carried out in this study, which are further detailed below. First, *in vitro* HTS data were first organized from the ToxCast/Tox21 program, including hit-calls and activity concentrations, focusing on Tox21 assays to maximize chemical coverage (Fig. 2, **step A**). Activity concentrations (AC_{50}s) were converted into equivalent aqueous or

intracellular concentration estimates using a model of chemical disposition within a cell-based *in vitro* assay system (Fig. 2, **step B**). *In vivo* biological responses were obtained through the evaluation of liver transcriptomic data from rats exposed to the same chemicals that were tested in Tox21 assays, organized in the DrugMatrix and TG-Gates databases (Fig. 2, **step C**). Pathway enrichment analyses, using gene ontologies available through four different databases, were carried out on the gene sets identified as differentially expressed in response to each exposure condition. High-throughput toxicokinetic (HTTK) modeling was carried out to convert *in vivo* dosing parameters into tissue-level concentration estimates (Fig. 2, **step D**). These data were then combined with physicochemical properties to develop predictive models of *in vivo* pathway-level alterations in the rat liver (Fig. 2, **step E**). These models predicted *in vivo* activity in 735 pathways for 221 chemicals, using *in vitro* bioactivity information from 144 Tox21 assays. Model performance was compared across different combinations of predictor/outcome variables, in order to compare and contrast the influence of including vs. excluding chemical disposition and toxicokinetic modeling (Table 1), as further detailed below.

2.2. Organization of in vitro Tox21 bioactivity data

HTS data were obtained through the ToxCast/Tox21 consortium using the database released in May 2019 (noted as 'invitroDB_v3.2') through the U.S. EPA web portal [13]. Tox21 assay bioactivity results were characterized using the following summary-level values for each chemical/assay pair: the winning (best-fit) concentration–response model as determined by the ToxCast pipeline (constant, Hill, or gain/loss); fitted AC_{50} value (concentration at which the activity reaches 50% of its maximal values for a chemical/assay pair); fitted Top value (maximum possible response predicted by the winning concentration–response model); maximum median response value (max_med); and activity hit-call. Hit-calls represent overall assay endpoint activity [24–26]. Briefly, if a chemical was tested in multiple-concentration format in an assay, with sufficient concentration–response data to fit a model, then hit-call is 1 (active) if the winning concentration–response model was not constant and if both Top and max_med exceed an assay-specific activity cutoff value; otherwise, hit-call is 0 (inactive) [25,26]. Note that at this stage, fitted AC_{50} values were retained for all chemical/assay pairs with non-constant concentration–response model, even if they had hit-call 0. This differs from the usual ToxCast convention to substitute a placeholder value of 1e6 μM for AC_{50} whenever hit-call is 0. (In terms of variables defined in the invitroDB v3.2 README, we used "modl_ga", not "ac50".) For chemical-assay pairs where the winning concentration–response model was constant, invitroDB_v3.2 reports NA for fitted AC_{50} and Top values. In these cases, the following placeholder values were substituted: AC_{50} = 1000 μM (the highest concentration tested), and Top value = 0 (because a constant concentration–response model represents a constant response level of 0). Potential *in vitro* assay interference from cytotoxicity and cell stress was also considered [25]. At cytotoxic concentrations, a "burst" of activity is seen across a characteristic set of assays, classified as "cytotoxicity-burst assays" [25,27].

Cytotoxicity-burst assays were identified and retained for calculation of cytotoxicity points, described in a later section.

2.3. Modeling chemical disposition within the in vitro Tox21 assay systems

In vitro activity concentrations (AC_{50}) for each chemical/assay pair were converted into estimates of equivalent aqueous and intracellular concentrations using the Armitage model, an *in vitro* chemical disposition model implemented in the "httk" R package, version 2.0.1 [21,28]. The Armitage model, uses physicochemical properties and assay-specific information including plate and well geometry, cell yield (number of cells in each well), and the fraction of fetal bovine serum in the reagent (if any). Assay-specific information was gathered from Tox21 assay documentation [29], and included the plate type (clear flat bottom or solid flat bottom 1536-well plates) and fraction of fetal bovine serum in the reagent. All 144 included Tox21 assays were confirmed to be cell-based assays (not cell-free assays). Default cell yield assumptions were used, originally based upon Corning technical documentation [30]. Well geometry for each plate type was used as built into the "httk" package implementation. The relevant assay-specific information is included in supplementary material (**S1 File**).

2.4. Determination of pathway-level activities within the rat liver using transcriptomic data from DrugMatrix and TG-Gates

Biological responses in the rat were evaluated using transcriptomic data obtained from two publicly available resources: DrugMatrix and TG-Gates. Array data across both databases were organized and analyzed using similar methods to maximize consistency. Data from the DrugMatrix database were collected as previously described [31]. The majority of data represented profiles collected 24-h post-exposure, though when unavailable, data collected proceeding daily doses lasting 3- or 5-d were used. Dosing regimens for the included experiments largely consisted of oral gavage administrations, with some intraperitoneal, intravenous, and subcutaneous dosings. Chemicals were administered daily at 1–3 doses in addition to the vehicle control. Vehicle controls consisted of either 0.5% carboxymethyl cellulose, corn oil, saline, or water. Transcriptomic data from the DrugMatrix database were generated using liver RNA samples hybridized to the Affymetrix GeneChip Rat Genome 230 2.0 array. Data were processed and analyzed by the NTP using established methods for normalization, QA/QC, and statistical assessment, as previously summarized [23,31]. Statistical significance was evaluated using the t-statistic with an Empirical Bayes method of estimating variance, as previously detailed [32].

Data from the TG-Gates database were collected as previously described [33]. For the current project, transcriptomic data included Affymetrix GeneChip Rat Genome 230 2.0 array data collected from rats exposed via single dose to a chemical, with liver samples collected 24 h post-exposure. Similar to the DrugMatrix experiments, dosings largely consisted of oral gavage administrations, with some intraperitoneal, intravenous, and subcutaneous dosings. Chemicals were administered daily at 1–3 doses in addition to the vehicle control. Vehicle controls consisted of either 0.5% carboxymethyl cellulose, corn oil, saline, or water. Transcriptomic data from TG-Gates were evaluated for QA/QC metrics, as previously described [33]. In the current analysis, array data were downloaded from the online TG-Gates database [34] and processed and analyzed using the Linear Models for Microarray Data (LIMMA) package [35] in R Software, v3.5.2. Data were normalized using robust multi-chip average and evaluated for differential expression comparing gene

expression levels from tissues of treatment groups vs. control groups (three animals per group). Significance was established using the t-statistic with an Empirical Bayes method of estimating variance, paralleling the DrugMatrix data analysis.

Probesets were identified as significantly differentially expressed when meeting a fold change ≥ ±1.5 (average exposed/average control) and $p < 0.01$ (exposed vs. control), paralleling our previous investigation evaluating *in vitro*-to-*in vivo* response concordance with the DrugMatrix database [23]. This represents a relatively relaxed filter to maximize the detection of potential gene expression changes leading to pathway-level activities. Probeset annotation information was updated to reflect the most recent rat genome annotation release through Affymetrix (v36). Probesets meeting statistical significance criteria were referred to as differentially expressed probes (DEPs) and were used to filter for unique lists of differentially expressed genes (DEGs).

Biological pathway-level activities associated with each chemical exposure were determined through enrichment analyses of the DEGs. Pathway annotations were organized from several databases to evaluate whether pathway-level predictions varied across knowledgebase. Specifically, six different pathway databases were used, four of which represented gene sets extracted from the Molecular Signatures Database (v7.0): BIOCARTA, KEGG, PID, and REACTOME [36,37]. Canonical pathways were also analyzed from Ingenuity Pathway Analysis (IPA), derived from the Ingenuity Systems® Knowledgebase. One gene set was also included from the Molecular Signatures Database, namely the HALLMARK gene set, and is referred to under the umbrella term of pathway database in this study to allow for concise descriptions. Pathways were filtered for those that included a minimum of 10 rat genes. Pathways were analyzed for significant enrichment amongst lists of DEGs for each chemical exposure using the R package, Platform for integrative analysis of omics data (PIANO) (v3.4.1) [38], with pathways identified as significant using the Fisher's Exact test p-value < 0.05. This filter was implemented as it parallels previous pathway-level investigations evaluating chemical-induced toxicity [39–46] and was shown to result in greater *in vitro-in vivo* concordance in our previous analysis in comparison to using a multiple test corrected p-value filter [23]. Pathways that were identified as significant were considered 'active', and pathways that were not identified as significant were considered 'inactive'.

2.5. Toxicokinetic modeling to estimate tissue-level concentrations resulting from in vivo dosings

The doses administered to rats in DrugMatrix or TG-Gates experiments were converted to estimates of tissue concentrations using toxicokinetic modeling. A generic physiologically-based toxicokinetic (PBTK) model was used, as implemented in the U.S. EPA's "httk" R package (version 2.0.1, model "pbtk") [12,47]. This model was used to predict the time course of liver and plasma concentrations (μM) in the rat during each *in vivo* experiment, based on information surrounding the chemical, daily dose, duration of dosing, and route of administration in each experiment. The PBTK model also required chemical-specific data regarding intrinsic hepatic clearance rate and the fraction of the chemical unbound in plasma protein, as well as physical–chemical properties used to predict tissue partitioning. The

model can predict time-dependent concentrations in many compartments, though liver and plasma were selected as the most relevant tissues here. To summarize the time-dependent concentrations, maximum and mean concentrations (μM) over the duration of each experiment were computed, for both rat liver and rat plasma. Together, four different TK-predicted *in vivo* body concentrations (C_{vivo}) were calculated for each experiment: liver maximum, liver mean, plasma maximum, and plasma mean.

2.6. Organization of physicochemical descriptors

Physicochemical descriptors were included as predictor variables, as we recently demonstrated that including physicochemical properties improves the overall concordance between *in vitro* and *in vivo* responses [23]. Physicochemical data were obtained through the U.S. EPA's Computational Toxicology Dashboard [48]. In instances when more than one experimental value was presented, the average across the available values was used. When experimentally derived data were not available, predicted values were used. The following set of physico-chemical descriptors was used: boiling point, $\log_{10}$ Henry's Law constant, melting point, $\log_{10}$ molecular weight, $\log_{10}$ octanol–air partition coefficient (logKoa), $\log_{10}$ octanol–water partition coefficient (logP), $\log_{10}$ vapor pressure, and $\log_{10}$ water solubility. After the final selection of included chemicals (see following section), each physicochemical descriptor was centered and scaled to have mean 0 and standard deviation 1 across the included chemicals. The center and scale values for each physicochemical descriptor are included in supplementary material (**S2 File**).

2.7. Selection of included chemicals and Tox21 assays

First, chemicals were filtered to include only those that satisfied the following data requirements: Chemicals were required to have liver transcriptomic data within the DrugMatrix and/or TG-Gates databases from rats acutely exposed (1–5 daily dosings). Additionally, chemicals were required to have sufficient TK information to perform TK modeling, where TK data were collected based on *in vitro* or *in vivo* measured TK data, or were predicted *in silico* based on values previously reported [49]. Lastly, chemicals were required to have sufficient physicochemical data available to run the chemical disposition model. After these selection criteria were applied to the list of chemicals, a further selection of the remaining chemicals and Tox21 assays was performed, to ensure that all retained chemicals were tested in multiple-concentration format in all retained Tox21 assays (i.e., all chemicals had Top, max_med, AC_{50}, and winning-model information reported for all assays).

Finally, the remaining Tox21 assays were filtered to retain only those that satisfied the following requirements: Assays were required to demonstrate activity (hit-call of 1) associated with at least one of the included chemicals. Additionally, assays were removed from consideration if they were labeled as "artifact detection" assays in ToxCast invitroDB v3.2 to remove redundant and inherently highly-correlated measures. A table of assays with the reason for inclusion or removal is included in supplementary material (**S3 File**). After this final selection step, 221 chemicals and 144 Tox21 assays were retained.

2.8. Constructing training and test sets and mitigating data imbalance in predictive models

To assess whether model performance is generalizable (i.e., not a result of overfitting), three-fold cross validation was performed. Data were randomly divided into three equal-sized folds. In each round of cross-validation, one fold (1/3 of the data) was selected as the test set, and the other two folds (2/3 of the data) were combined and used as the training set. Machine-learning models were trained using the training set, and then their performance was evaluated using the test set.

Because most experiments were associated with inactivity across most biological pathways, the training set for each pathway was usually imbalanced, with higher incidence of inactivity vs. activity. Highly imbalanced data can present difficulties for machine-learning classification approaches. To mitigate these difficulties, for each training set that had fewer than 30% of experiments active in the specified pathway, we improved balance by applying the Synthetic Minority Over-Sampling Technique (SMOTE) [50]. Briefly, SMOTE randomly synthesizes additional data items (rows) in the minority class using k-nearest-neighbors (KNN) interpolation, and randomly downsamples data items in the majority class, to yield a more-balanced training set.

Here, SMOTE was performed using the following set of features: scaled physical–chemical properties; the four TK-predicted $\log_{10}$ C_{vivo} descriptors; *in vivo* $\log_{10}$ dose; and continuous Tox21 chemical-assay response metrics: max_med responses, Top values, and nominal $\log_{10}$ AC_{50} values. These continuous Tox21 chemical-assay response metrics were used for KNN interpolation, rather than using binary hit-calls, because hit-calls are a dichotomized summary of the continuous metrics. Interpolating based on the underlying continuous metrics, then dichotomizing (hit-calling), yields a better representation of the distribution of assay responses compared to interpolating based on the pre-dichotomized metrics (hit-calls).

Note that only nominal $\log_{10}$ AC_{50} values were used for SMOTE, rather than also including the Armitage-model-predicted intracellular and aqueous-phase $\log_{10}$ AC_{50} values. This choice was made because intracellular and aqueous-phase $\log_{10}$ AC_{50} values are simply functions of nominal $\log_{10}$ AC_{50} values and physicochemical properties, and these variables would therefore be highly correlated, which might skew the identification of nearest-neighbors for KNN interpolation. Therefore, SMOTE was performed using only nominal $\log_{10}$ AC_{50} values, and then the Armitage model was applied after SMOTE.

SMOTE was applied to the training set for each pathway. The number of synthesized minority-class items and sampled majority-class items varied, depending on the degree of imbalance for each pathway. Therefore, after SMOTE was applied, each pathway had a slightly different training set. Synthesized data items are interpolated, so they do not correspond to any actual chemical or *in vivo* experiment; instead, they represent hypothetical “chemicals” and “experiments”. SMOTE was not applied to the test sets. Training sets before applying SMOTE; training sets after applying SMOTE; and test sets are all available in supplementary material (**S4 File**).

2.9. Hit-calling for synthesized data items in SMOTED training sets

For each data item (row representing a real or hypothetical chemical) in each training and test set, assay hit-calls were made based on the following criteria: If both Top and maximum median response exceed an assay-specific activity cutoff level (as reported in the invitroDB_v3.2 database), then hit-call is 1 (active). Otherwise, hit-call is 0 (inactive). Note that the default hit calling within invitroDB_v3.2 also requires that the winning model not be constant. Here, that criterion is effectively enforced by the criterion that Top > assay-specific activity cutoff, because for all cases where the winning model was constant, Top was fixed at zero, and is therefore below the assay-specific activity cutoff. For all data items that were not synthesized by SMOTE, these hit-calls were identical to the existing Tox21 hit-calls. (Data items synthesized by SMOTE did not have existing Tox21 hit-calls.) After hit-calling was performed, for each data item, $\log_{10}$ AC_{50} values for all assays with hit-call 0 were replaced with a placeholder $\log_{10}$ AC_{50} value of 99. In this way, $\log_{10}$ AC_{50} = 99 in a given assay serves as a flag for inactivity in that assay.

3. Applying chemical-disposition models to synthesized data items in SMOTED training sets

For each data item in each training and test set, the Armitage chemical-disposition model was applied to the nominal $\log_{10}$ AC_{50} values for each assay, using the physical–chemical properties for the data item, and the assay-specific properties (e.g. well geometry, fraction of fetal bovine serum, cell yield) for each assay. The result was a corresponding intracellular $\log_{10}$ AC_{50} value and aqueous $\log_{10}$ AC_{50} values for each assay and each data item. If hit-call was 0 and therefore nominal $\log_{10}$ AC_{50} had been assigned a placeholder value of 99, then both intracellular and aqueous $\log_{10}$ AC_{50} values were also assigned a placeholder value of 99.

3.1. Calculating in vitro cytotoxicity concentrations

Potential *in vitro* assay interference from cytotoxicity and cell stress was considered (Judson et al. 2016). At cytotoxic concentrations, a “burst” of activity is seen across many assays, classified as “cytotoxicity-burst assays” (EPA 2019e; Judson et al. 2016). A summary “cytotoxicity point” for each chemical is commonly computed as the median $\log_{10}$ AC_{50} value across cytotoxicity-burst assays. Although invitroDB3.2 includes pre-calculated cytotoxicity points for each chemical, we re-calculated these values in order to develop cytotoxicity points based on the cytotoxicity-burst assays included in the retained set of 144 Tox21 assays, and to develop cytotoxicity points based on intracellular and aqueous-phase $\log_{10}$ AC_{50} estimates predicted using the Armitage chemical-disposition model. Specifically, cytotoxicity points were calculated as the median of AC_{50}s across the cytotoxicity-burst Tox21 assays, using either nominal AC_{50}s or estimated intracellular or aqueous-phase equivalent AC_{50}s. If any chemical did not have a hit-call of 1 in at least two cytotoxicity-burst assays (so that a median could not be calculated), it was assigned the following default cytotoxicity point (in units of $\log_{10}$ μM): 3 for nominal concentrations; 3 for aqueous-phase concentration estimates; and 5 for intracellular concentration estimates. These default cytotoxicity points reflected the highest-observed or estimated AC_{50} among active assays.

3.2. Random forest modeling to evaluate predictivity of Tox21 bioactivity against in vivo biological responses

Random forest modeling was used to evaluate the potential impact of toxicokinetic modeling on the overall accuracy of using Tox21 bioactivity to predict *in vivo* pathway-level activity in the rat liver. Briefly, random forest modeling builds an ensemble of decision-tree models [51]. Each tree is trained on a bootstrap resampling of the training set ("in-bag" samples); approximately one-third of data items are left out for each tree ("out-of-bag" samples). Within each tree, each split is chosen from a randomly-selected subset of the predictor variables (a different random selection is made for each split). Each tree in the ensemble then "votes" on the ultimate classification of each item. The final result is the fraction of trees voting for each category. Here, random forest modeling was implemented using the "randomForest" R package [52]. Five thousand trees were included in each random forest. For each pathway that was evaluated, 12 separate random forest classifier models were trained and tested (Table 1). Models were built using various sets of predictor variables that either included or excluded TK-predicted C_{vivo}. This design allowed for the direct comparison of the predictive ability of models with and without TK. Note that, unlike our previous analysis [23], TK information was not used as a dose-applicability filter for DrugMatrix/TG-Gates experiments. Instead, the TK-predicted C_{vivo} values were used directly as predictor variables.

To evaluate whether the predictive ability of these 12 models could have occurred by chance, 12 additional random-forest classifier models were also trained using randomly-permuted versions of each set of predictor variables from the training set, and tested using the original test set, yielding a total of 24 models in this study. The performance of each model could then be compared to the performance of its permuted version. The predictions of each of the 24 random forest models for each pathway are included in the supplementary material (**S5 File**).

3.3. Predictive model performance evaluation and results ranking

The performance of each model for each test set (in three-fold cross-validation) was assessed by calculating the area under the receiver-operator characteristic curve (AUC-ROC). Briefly, for a range of thresholds on the fraction of "active" votes needed for the "active" classification to win, the ROC is the curve traced out by plotting the true positive rate vs. the false positive rate for each threshold. The area under the ROC is a metric of model performance: AUC above 0.5 indicates performance better than chance, and AUC below 0.5 indicates performance worse than chance. For each pathway and each fold of cross-validation, AUC-ROCs were computed for each of the 24 models using R package "pROC" [53]. Additionally, the performance of two models can be evaluated by comparing their AUC-ROCs, yielding a p-value measuring whether one model performs significantly better than another. For each pathway and for each fold of cross-validation, the AUC-ROC of each model was compared to the AUC-ROC of the corresponding permuted version of that model, to determine whether each model outperformed random noise at a statistically-significant level. The AUC-ROC comparisons were performed using the method of Delong et al. (1988) as implemented in R package "pROC" [53]. The AUC-ROC comparisons and their p-values are included in supplementary material (**S6 File**).

For each fold of cross-validation and for each pathway, models were ranked from highest to lowest AUC-ROC. The model with the highest AUC was considered the "winning" model for that pathway and that fold. Then, within each fold of cross-validation, the number of "wins" for each model was counted (i.e., the number of pathways where each model had the highest AUC-ROC). Models were then ranked from highest to lowest total number of "wins". To assess how model performance varied across pathway databases, model ranking was also repeated separately for the six different pathway databases (i.e., BIOCARTA, KEGG, IPA, PID, and REACTOME).

A final version of the highest-ranked random forest model (Model 10 in Table 1) was then trained using all of the data, rather than the 2/3 training set previously employed. The SMOTED training sets for this final version of the model are included in supplementary material (**S7 File**); the resulting random forest model objects themselves are also included in supplementary material (**S8 File**). This final version of the model was used for the rest of the analyses.

3.4. Interpreting models through variable importance measures and feature contribution plots

Visualizations were generated with the goal of gaining insight into potential relationships between predictor variables and response variables included in the most highly-ranked model: (1) predictor-variable importance rankings from the final version of the winning random forest model; and (2) feature contribution plots from the final version of the winning random forest model. For the variable-importance plots, importance was calculated as the mean decrease in accuracy in predicting out-of-bag samples when each predictor variable was permuted. This metric assesses the total importance of each predictor, both on its own and interacting with other predictors. To summarize variable importance across pathways, the mean decrease in accuracy for each variable was averaged across all pathways; this "grand mean" decrease in accuracy was then used to rank the average importance of each variable across pathways. For the feature contribution plots, the overall contribution of each predictor variable towards the probability of "active" classification for each pathway was calculated, and resulting feature contribution plots were produced for the most important variables using R package "forestFloor" [54].

3.5. Leveraging parameterized models to predict in vivo activity across all chemicals in Tox21

As an example of the utility of the resulting models, we applied the winning random forest model to predict *in vitro* pathway-level activity for 6617 chemicals tested in Tox21 that had sufficient data coverage. These chemicals represented those that were remaining within the database that had been tested in multiple-concentration format across the 144 included assays and had complete physico-chemical property data available (6711 chemicals) and had available TK parameters (6970 chemicals); the intersection of these two sets of chemicals comprised 6617 chemicals. Armitage-converted Tox21 assay AC_{50}s, physicochemical properties, and TK predictions for *in vivo* concentrations were gathered using the previously detailed methods. Doses tested for this exercise included the 5th, 50th, and 95th percentiles on the log scale of the doses tested in DrugMatrix and TG-Gates via oral gavage (reflecting

the most commonly evaluated exposure route). TK predicted *in vivo* concentrations for each of the 6617 chemicals at each of the three doses are included in supplementary material (**S9 File**); Tox21 assay AC_{50}s, physicochemical descriptors, and ExpoCast predicted exposures are included separately in the supplementary material (**S10 and S11 Files**). Results from this modeling effort presented predictions of whether or not pathways would be altered upon exposure to chemicals in Tox21. The resulting predictions are included in the supplementary material (**S12, S13,** and **S14 Files**).

4. Results

4.1. Overview of data organization

Several data requirements were used to first filter chemicals and Tox21 assays, as detailed in the Methods. After these filters, data were retained for 221 chemicals and 144 Tox21 assays. These 221 chemicals were evaluated across a total of 519 DrugMatrix and TG-Gates experiments. *In vivo* pathway activity was originally assessed in 2538 pathways across six different pathway databases. Because most of these pathways were inactive in most or all of the 519 *in vivo* experiments, pathways were further filtered to include only those that were active in at least 10% of experiments, leaving 735 pathways for analysis. This requirement ensured that for each pathway, at least a few experiments were active in each training and test set.

These filtered data were then prepared for model building. Specifically, data were merged to produce a wide-format table whose rows represented the *in vivo* experiments (identified by dataset [DrugMatrix or TG-Gates], chemical, dose, duration, route, and vehicle of administration) and whose columns were the following predictor variables that were tested in various combinations: physical–chemical properties (centered and scaled to standardize each physical–chemical property to have a mean of 0 and a standard deviation of 1 across the chemicals); Tox21 assay max_med values; Tox21 assay Top values; Tox21 assay $\log_{10}$ AC_{50}s; and the TK-predicted body concentrations: maximum liver concentration, maximum plasma concentration, mean liver concentration, and mean plasma concentration. These predictor variables are fully provided as supplementary material (**S10 File**). In addition to these predictor variables, each row of the table also had a corresponding activity call for each of 735 pathways (i.e., 735 separate response variables). These response variables are provided as supplementary material (**S15 File**).

For building and testing predictive models, *in vivo* experiments were randomly divided into three folds (**S15 File**), and training and test sets were formed for each pathway. To address data imbalance resulting from a high prevalence of *in vivo* response inactivity, SMOTE was applied to training sets (but not test sets) for each pathway and each fold of cross-validation. SMOTE synthesizes new “active” experiments using k-nearest-neighbor interpolation to better explore regions of predictor variable values associated with activity. Then, for SMOTEd training and non-SMOTEd test sets, hit-calls were made, intracellular and aqueous AC_{50} estimate values were computed, and cytotoxicity points were calculated across Tox21 data, as detailed in the Methods. These training and test sets are included in supplementary material (**S4 File**). These data were then used to build and test predictive models.

4.2. Including toxicokinetic modeling improved predictive performance

Predictive models were built and tested using various sets of predictor variables (Table 1), all aimed at using *in vitro* Tox21 data to predict pathway-level alterations in the rat liver. Model performance was ranked based on the number of pathways for which each model had the highest AUC-ROC. Models that included TK outranked models that did not include TK (Fig. 3). The highest-ranking model included Tox21 AC_{50} values converted to aqueous (media) concentrations, and *in vivo* doses converted to TK-predicted circulating plasma concentrations (i.e., 'Model 10′ in Table 1). Model performance rankings were compared across different pathway databases, where the REACTOME database showed the most consistent model performance rankings in comparison to all the databases combined, as it contained the largest number of pathways (supplementary material, **S16 File**). All random forest results, including fraction of classification trees voting "active" for each experiment for each combination of parameters across the training and test sets are available through Dataverse [55]. An example ROC curve is shown in Fig. 4 for the first-fold training set. All AUC-ROC values for each model and each pathway (for training and test sets for each fold of cross-validation) are provided in supplementary material (**S17 File**); ROC plots for all pathways for the highest-ranking model are provided in supplementary material (**S18 File**).

4.3. Biological interpretation of features contributing to the random forest model predictions

Unlike regression models, random forest models do not explicitly describe the nature of the inferred relationships between the predictor variables and the response variable [56]. For example, does an increase in TK model-predicted maximum plasma concentration correspond to an increase or decrease in pathway activity probability? How steep is the relationship? Is it modified by interactions with other predictor variables? To shed light on the nature of the model-inferred relationships between predictor and response variables for each pathway, the following analyses were performed for the models for each pathway: (1) variable importance rankings; and (2) feature contribution plots. These were carried out using results from the highest-ranking model (Model 10 in Table 1), re-trained using all data rather than the 2/3 training set for cross-validation. Such analyses represent qualitative first steps towards investigating the biological basis of resulting predictive models.

4.4. Variable importance rankings from the highest-ranking random forest model for each pathway

Predictor-variable importance rankings from the highest-ranking model are illustrated here for an example pathway, the PXR-RXR activation pathway (Fig. 5A), with variable importance rankings for the highest-ranking model for all pathways provided in supplementary material (**S19 File**). Variable importance here is measured by the mean decrease in accuracy in predicting out-of-bag samples when each variable is permuted; variable importance is ranked from highest to lowest mean decrease in accuracy. This is a standard metric for assessing variable importance in random forest models [51]. The PXR-RXR pathway was used as an example in Fig. 5A because it represents a common pathway involved in liver toxicity, with pertinence in drug testing and chemical safety evaluations [57]. For this pathway, dose of the chemical administered ($\log_{10}$ dose) is the most important

predictor, followed by the physicochemical parameter, logP, and then concentration of the chemical estimated to occur within the *in vivo* tissue ($\log_{10} C_{vivo}$ in plasma). The top three Tox21 assays included two cell viability assays and one androgen receptor (AR) agonism assay. Other top-ranking assays include assays for sonic-hedgehog signaling (SHH) pathway agonism, other AR agonism and antagonism assays, progesterone-receptor (PR) antagonism assays, and constitutive androstane receptor (CAR) agonism assays. Interestingly, of the two assays which might be expected *a priori* to predict PXR-RXR activation — the TOX21 PXR agonism assay and the Tox21 RXR agonism assay — neither one ranks in the top 30 most important variables when predicting PXR-RXR activation *in vivo*.

This result likely occurs because there was much less *in vitro* activity in the PXR/RXR agonism assays than there was *in vivo* activity in the PXR-RXR pathway: activity in the *in vitro* RXR and PXR assays occurred in only 8 and 9 chemicals (corresponding to 19 and 24 experiments) included in this analysis, respectively, whereas 310 experiments exhibited activity in the *in vivo* PXR-RXR pathway. This difference between *in vitro* and *in vivo* activity prevalence is likely caused by the much higher range of *in vivo* internal doses compared with the range of *in vitro* tested concentrations: the maximum tested *in vitro* concentration for the included chemicals was 100 μM, whereas 100 μM is only the 20th percentile of *in vivo* internal peak plasma concentrations for the included experiments. In other words, if higher concentrations had been tested *in vitro*, more chemicals might have shown activity in the PXR and RXR agonism assays, and that *in vitro* activity might have been predictive of *in vivo* activity in the PXR-RXR pathway. However, given the available range of tested *in vitro* concentrations, activity in other assays — in which more chemicals show activity at concentrations within the tested range — is identified by the machine-learning model to be predictive of *in vivo* PXR-RXR pathway activity.

A global measure of variable importance, across all included pathways, is summarized in Fig. 5B, a heatmap of the mean decrease in accuracy in predicting each pathway (rows) when each variable (columns) was perturbed. Findings show that globally across pathways, the most important variables for predicting pathway-level activity are HTTK-predicted maximum *in vivo* plasma concentration; administered *in vivo* dose; and physicochemical parameters. Variables representing *in vitro* assay AC_{50}s exhibit lower average importance across all pathways, although certain assay AC_{50}s are still important for predicting activity in certain pathways.

The variable importance measure shown in Fig. 5A and B does not explain the direction of the relationship between each predictor and the response; it also does not explain whether the relationship is significant for that predictor alone, or whether interactions between that predictor and other predictors are important. However, variable importance rankings do provide important information about the overall contribution of each variable to model performance. Together, these findings demonstrate that metrics of external and internal dose are important, along with bioactive concentrations in assays that may not be *a priori* hypothesized to contribute towards *in vivo* pathway activation in response to chemical exposures.

4.5. Feature contribution plots from the highest-ranking random forest model for each pathway

Feature contribution plots for all pathways (for the highest-ranking random forest model) are provided in supplementary material (**S19 File**), and an example plot is included here also focusing on the PXR-RXR activation pathway (Fig. 6). Feature contributions can be used to understand the change in the overall probability that the random forest model predicts activity that is attributable to a given predictor variable at a given value [54]. For example, the relationship between PXR-RXR pathway activity and the concentration estimated to elicit PXR-RXR pathway activity ($\log_{10} C_{vivo}$) in the rat plasma (third panel from left in the top row of Fig. 6) appears to be sigmoidal in shape, with $\log_{10} C_{vivo} < 2$ corresponding to a nearly-constant decreased probability of activity and $\log_{10} C_{vivo} > 2$ corresponding to an increasing probability of activity. This result suggests a relationship that may be based on a threshold value of the plasma concentration around 100 μM. The relationship between PXR-RXR pathway activity and logP (second panel from left in the top row of Fig. 6) indicates that the probability of pathway-level activity *in vivo* decreases dramatically as logP of the chemical decreases below approximately 1, and is increased by an approximately constant amount for logP above 1, also indicating a potential threshold in the relationship. For the *in vitro* Tox21 AC_{50}s predictor variables, cell viability assays (e.g., TOX21_RT_HEK293_FLO_08hr_viability) show the general trend where *in vivo* PXR-RXR pathway activation probability decreases as viability assay AC_{50} decreases (representing increased *in vitro* potency). Conversely, other *in vitro* Tox21 predictor variables (e.g., AR agonist, AR antagonist, SHH agonist, CAR agonist, and PR antagonist assays) show relationships to increased PXR-RXR pathway activation probability alongside decreased AC_{50} values (Fig. 6). Feature contribution plots can also be color-coded by a specified predictor variable, allowing for the qualitative assessment of potential interactions with a predictor variable of interest. Fig. 6 is color-coded by the $\log_{10}$ *in vivo* dose. Interactions are not blatantly obvious, suggesting that the illustrated other predictor variables may act mostly independent of log10 dose in this specific example.

4.6. Assessing contribution of Tox21 predictor variables

All compared models in Table 1 include Tox21 predictor variables (hit calls or AC_{50}s). To assess whether including Tox21 predictor variables improves model performance vs. not including these predictors, an additional set of models could be trained that excluded all Tox21 predictors, but retained various combinations of dose/physicochemical and toxicokinetic predictors. However, such a model performance comparison would be complicated by the substantially-smaller number of predictor variables for non-Tox21 models: because the random forest algorithm selects the best split from a randomly-selected subset of predictors at each split of each tree, a model with a large number of predictors may perform differently than a model that includes a small subset of those predictors. In other words, the number of predictor variables included in the model can significantly influence model performance and therefore minimize the relevance of such a comparison. To assess the contribution of Tox21 variables while controlling for the potential influence from the predictor variable number, the top 20 most important variables were identified for each pathway from the highest-ranked model identified in the foregoing analysis. The model was

re-fit (in three-fold cross validation) with only those top 20 variables, called the “top 20” model. Then, any Tox21 variables were excluded from the top 20, and the model was again re-fit (in three-fold CV) with that non-Tox21 subset of the top 20 variables (the exact number of variables depended upon how many Tox21 variables were in the original top 20), called the “top without Tox21” model. The AUC-ROC was computed for the test set for each pathway and each fold of cross-validation, for both the “top 20” model and the “top without Tox21” model, and then averaged across the three folds of cross-validation. The average test set AUC-ROC was greater for the “top 20” model than for the “top without Tox21” model in 543 out of 735 pathways, indicating that Tox21 predictor variables do contribute important information for predicting pathway-level activity.

4.7. Prediction results across all biological pathways for chemicals in Tox21 using the winning predictive model

The utility of the highest-performing model (‘Model 10’ in Table 1) was demonstrated by generating new predictions of *in vivo* biological responses across all biological pathways for chemicals tested in Tox21 with sufficient data coverage (n = 6617). The random forest model predicts the probability of activity at a specified dose, not the dose at which activity occurs. Therefore, the model was evaluated to make activity predictions at a set of three theoretical doses, spanning the doses evaluated in the DrugMatrix/TG-Gates databases. Doses were selected as the 5th, 50th, and 95th percentiles of the doses tested in DrugMatrix and TG-Gates. These doses correspond to 2.4 mg/kg/day (5th percentile); 150 mg/kg/day (50th percentile); and 2000 mg/kg/day (95th percentile). It is notable that these doses are all fairly high in the context of environmental relevance; the median ExpoCast predicted exposure for these 6617 chemicals in humans is approximately 5e-6 mg/kg/day, and the highest ExpoCast predicted exposure for these 6617 chemicals is approximately 5 mg/kg/day, on the order of the 5th percentile dose. ExpoCast predicted exposures for these chemicals are available in the supplementary material (**S11 File**)

The resulting probability of activity for each chemical in each pathway was visualized using heat maps, focusing on pathways with the highest predictive performance (Fig. 7). These top-ranking pathways specifically included those with the top 10% AUC-ROC for the winning model, with AUCs ranging from about 0.84 to about 0.76. Model-predicted probabilities of activity for all pathways are available in the supplementary material (for the 5th percentile dose in **S12 File**; for the 50th percentile dose in **S13 File**; and for the 95th percentile dose in **S14 File**). Three key points are clear from these heatmap visualizations. First, the pathways with the overall highest AUC-ROC, as shown in Fig. 7, include those of high relevance to the liver, including many pathways involved in metabolism, as well as inflammation/immune response, cell growth, cell differentiation, and cell death, among others. Second, biological activity generally increases with dose for most chemicals: there is almost no activity in any pathway at a dose of 2.4 mg/kg/day; more activity at 150 mg/kg/day; and even more activity at 2000 mg/kg/day. Third, there are some chemicals that exhibit activity probability $>50\%$ broadly across pathways at a dose of 150 mg/kg/day (top half of the heatmap), and others that exhibit more selective pathway activity at the same dose (bottom half of the heatmap). A rough difference between these two groups of chemicals is revealed by annotating the heatmap with the plasma concentration (C_{vivo}): the chemicals

with more selective pathway activation have generally lower plasma concentration at the same dose than the chemicals with less selective activation. This result reflects the importance of TK in the best-performing model. However, these differences do not fully explain the large differences in pathway selectivity. Moreover, the activity predictions could be evaluated for these 6617 chemicals by screening the literature to compile evidence for potential liver disruptions at the evaluated doses. It would be interesting in future studies to evaluate additional factors contributing to pathway selectivity and additional evidence streams relating to liver toxicity.

5. Discussion

There is an ever-expanding need to improve computational models that can better leverage findings from *in vitro* systems to predict *in vivo* biological responses resulting from chemical exposures and/or pharmaceutical treatments. This study serves as a critical advancement towards this effort by demonstrating how recent developments in chemical disposition and toxicokinetic modeling can be used to improve model performance in predictive biology applications. Here, Tox21 HTS data were used to train and test several models, all aimed at predicting pathway-level activities derived through transcriptomic evaluation of the livers of rats exposed acutely to one of 221 chemicals. Methods addressed certain limitations inherent in high-dimensional chemical screening data, including data imbalance caused by many inactive responses. Resulting computational models were able to successfully incorporate chemical disposition modeling and toxicokinetics, yielding improved predictive model performance. The two highest performing models were based on estimates in the *in vitro* system representing aqueous (media) and intracellular concentrations, and estimates in the *in vivo* system representing circulating plasma and liver concentrations.

Strategies were also incorporated to interpret biological relationships between modeled predictor and response variables, where unexpected combinations of multiple *in vitro* assays predicted *in vivo* pathway-level activities. Overall, this project provides a novel strategy through which *in vitro* data can be used to predict *in vivo* biological responses, which was lastly applied towards the prediction of *in vivo* responses across all chemicals in Tox21.

Our results demonstrate that the top-ranking predictive models incorporated chemical disposition estimates within *in vitro* cell-based systems. The majority of predictive toxicology studies using *in vitro* HTS data to predict *in vivo* outcomes have modelled *in vitro* activity and/or dose–response relationships using nominal concentrations (i.e., the concentration of a chemical(s) dissolved in solution and applied to the test system). There are inherent limitations to using nominal doses in such modeling efforts. To detail, the amount of chemical reaching and entering target cells within a test system may be non-linearly related to the nominal dose, as test chemicals can bind to extracellular components of *in vitro* systems in a non-linear manner [28,58,59]. Our results are in line with a recent study by Honda *et al.*, which found that *in vitro* activity concentrations converted through the Armitage model were better than nominal activity concentration at predicting *in vivo* points-of-departure based on apical toxicity [21]. Results are also congruent with a study by Casey *et al.* demonstrating that adjusting *in vitro* estrogen receptor activity concentrations

from nominal to intracellular concentrations substantially improved the ability to predict *in vivo* estrogenic activity [60]. These data, together, support the utility in modeling chemical disposition within *in vitro* cell-based systems within predictive biology applications.

Results consistently showed that the ability to predict *in vivo* biological responses was substantially improved by using toxicokinetic modeling. These results build upon our recent publication showing that *in vitro*-to-*in vivo* biological response concordance was significantly improved through the use of a TK filter for *in vivo* doses that are more comparable to *in vitro* activity concentrations [23]. The current study expands upon this approach by using forward TK to convert *in vivo* dosing information into a corresponding tissue concentrations and using those values as continuous predictor variables alongside *in vitro* activity information. The top performing model interestingly incorporated chemical concentration estimates circulating in plasma, while the second ranking model incorporated concentration estimates within liver tissue. These findings suggest that for future modeling purposes with similar uncertainties, plasma estimates may suffice in absence of tissue-specific data; though further research is needed to understand potential ranges of applicability. Together, results from this study provide unique evidence demonstrating the utility in incorporating chemical disposition and toxicokinetic modeling into predictive toxicology applications.

This analysis also employed methods to aid in the overall interpretation of the resulting models. These methods included the parsing of predictor-variable importance rankings from the random forest models and feature contribution plots. These additional views of the data all demonstrated that one *in vitro* assay does not simply inform whether or not the parallel molecular target will show activity *in vivo*. Previous studies have advocated for the use of orthogonal assays in informing *in vivo* biological responses [26,61]. Here, we expand on these findings by providing novel evidence supporting the utility of additional assays that may not be thought of as informative based on existing knowledge, though contribute valuable information towards informing and developing high performing computational models.

Mechanistic-based mappings of *in vitro* assays to *in vivo* pathways may not always identify the assays that are strongly correlated with *in vivo* activity. Rather, a “mechanistically-agonistic” approach that considers activity across assays, even if there is no obvious mechanistic connection with those assays, appears to do a better job of predicting *in vivo* activity. For example, we showed that assay response profiles from AR agonism and antagonism assays, progesterone-receptor (PR) antagonism assays, and constitutive androstane receptor (CAR) agonism assays are important predictor variables when predicting the likelihood of PXR-RXR activity in the rat liver. These are not necessarily assays that would be mapped *a priori* to *in vivo* PXR-RXR activation based on the biological understanding of PXR-RXR signaling. The importance of these assays in the random forest model does not necessarily imply mechanistic involvement of AR, PR, or CAR in PXR-RXR activation. The predictive power of these assays may well represent correlation, not causation. Our findings demonstrate that these correlations, though sometimes unexpected, provide useful information when predicting *in vivo* activity.

This study serves as an advancement in methods that can be used to interpret *in vitro* screening data to inform *in vivo* toxicity responses; though there remain additional steps that could further enhance this research area. One challenge faced in this analysis was the considerable class imbalance in data used to train and test predictive models. The majority of probed *in vitro* assays and *in vivo* pathways were inactive for the chemicals evaluated. This data imbalance presents a challenge for training a machine-learning model to predict activity vs. inactivity, since the model has relatively few examples of activity from which to learn. The problem of imbalanced data has been pervasive throughout research aimed at incorporating *in vitro* screening into predictive biology applications [14,23,62,63], and is a widely-recognized issue in many other applications of machine learning [64,65]. Here, we addressed this limitation through the application of an algorithm, namely, SMOTE, which was selected based on its ability to allow improved characterization of *in vitro* activity response distributions; though other approaches could be applied in future investigations. An additional challenge of this analysis was the high-dimensional feature space, with 144 Tox21 assays included in the analysis. Future investigations could benefit from feature selection methods to select the predictor variables that carry the most unique information. An additional challenge faced by predictive toxicology studies surrounds the availability of data spanning large numbers of chemicals to build and test *in silico* models. This study evaluated 221 chemicals that had adequate data coverage across the Tox21, DrugMatrix, and/or TG-Gates databases. As the generation of experimental data becomes increasingly higher throughput, it will be important to further develop and refine predictive models based on larger chemical domains. An additional challenge is the current paucity of high-throughput dose–response data for pathway enrichment analysis that would allow the derivation of a point of departure for pathway-level activity. This research gap will be addressed in part by future releases of high-throughput *in vitro* toxicogenomic screening data [66]. With point-of-departure data, models could be developed to directly predict the exposure level at which pathway-level activity might occur, rather than the probability of activity at a given exposure level. This research also could be applied towards understanding whether chemical disposition / toxicokinetic modeling aids in prediction of apical endpoints, as well as disease outcomes, in addition to the pathway-level responses evaluated here.

6. Conclusions

In conclusion, this study serves as an important advancement in predictive modeling by presenting approaches to more successfully leverage *in vitro* data to inform and predict *in vivo* biological responses to chemicals. These approaches are based on the careful consideration of *in vitro* chemical disposition, *in vivo* toxicokinetics, and machine learning methods. We specifically demonstrated that *in vitro* Tox21 data could be used to successfully predict *in vivo* biological responses in the rat liver, and leveraged the highest performing model to predict responses across all chemicals tested in Tox21. These approaches will undoubtably continue to expand in the upcoming years, resulting in increased confidence surrounding *in silico* and alternative test strategies to more rapidly identify chemical treatment strategies and evaluate the overall safety of chemicals in humans.

Funding

This work was supported by the American Chemistry Council Long Range Research Initiative and Foundation for Chemistry Research and Initiatives. Support was also provided through the National Institutes of Health (NIH) from the National Institute of Environmental Health Sciences, including grant funds (P42ES031007). Support was additionally provided through the Institute for Environmental Health Solutions at the University of North Carolina Gillings School of Global Public Health.

References

[1]. Agamah FE, Mazandu GK, Hassan R, Bope CD, Thomford NE, Ghansah A, Chimusa ER, Computational/in silico methods in drug target and lead prediction, Brief Bioinform. 21 (5) (2020) 1663–1675. [PubMed: 31711157]

[2]. Leelananda SP, Lindert S. Computational methods in drug discovery. Beilstein J Org Chem. 2016;12:2694–718. [PubMed: 28144341]

[3]. NAS. Toxicity Testing in the 21st Century: A Vision and A Strategy. Washington, DC: Committee on Toxicity Testing and Assessment of Environmental Agents, National Research Council, 2007 ISBN 978–0-309–10992-5.

[4]. NAS. Using 21st Century Science to Improve Risk-Related Evaluations. Washington, DC: Committee on Incorporating 21st Century Science into Risk-Based Evaluations; Board on Environmental Studies and Toxicology; Division on Earth and Life Studies; National Academies of Sciences, Engineering, and Medicine, 2017 ISBN 978–0-309–45348-6.

[5]. EPA US. Efforts to Reduce Animal Testing at EPA 2019 [cited 2020 Jan 3]. Available from: https://www.epa.gov/research/efforts-reduce-animal-testing-epa.

[6]. ECHA. Europoean Chemicals Agency (ECHA) Animal Testing under REACH 2020 [cited 2020 Jan 15]. Available from: https://echa.europa.eu/animal-testing-under-reach.

[7]. Wu Y, Wang G. Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. Int J Mol Sci. 2018;19(8).

[8]. Raies AB, Bajic VB In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev Comput Mol Sci. 2016;6(2): 147–72. [PubMed: 27066112]

[9]. Zavala J, Freedman AN, Szilagyi JT, Jaspers I, Wambaugh JF, Higuchi M, Rager JE New Approach Methods to Evaluate Health Risks of Air Pollutants: Critical Design Considerations for In Vitro Exposure Testing. Int J Environ Res Public Health. 2020;17(6).

[10]. Fry RC, Bangma J, Szilagyi J, Rager JE Developing novel in vitro methods for the risk assessment of developmental and placental toxicants in the environment. Toxicol Appl Pharmacol. 2019;378:114635. [PubMed: 31233757]

[11]. Wambaugh JF, Hughes MF, Ring CL, MacMillan DK, Ford J, Fennell TR, Black SR, Snyder RW, Sipes NS, Wetmore BA, Westerhout J, Setzer RW, Pearce RG, Simmons JE, Thomas RS, Evaluating in vitro-in vivo extrapolation of toxicokinetics, Toxicol. Sci 163 (1) (2018) 152–169. [PubMed: 29385628]

[12]. Ring CL, Pearce RG, Setzer RW, Wetmore BA, Wambaugh JF, Identifying populations sensitive to environmental chemicals by simulating toxicokinetic variability, Environ Int. 106 (2017) 105–118. [PubMed: 28628784]

[13]. EPA US. Exploring ToxCast Data: Downloadable Data 2019 [cited 2019 April 1]. Available from: https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data.

[14]. Liu J, Mansouri K, Judson RS, Martin MT, Hong H, Chen M, Xu X, Thomas RS, Shah I, Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure, Chem. Res. Toxicol 28 (4) (2015) 738–751. [PubMed: 25697799]

[15]. Thomas RS, Black MB, Li L, Healy E, Chu TM, Bao W, Andersen ME, Wolfinger RD, A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening, Toxicol. Sci 128 (2) (2012) 398–417. [PubMed: 22543276]

[16]. Cox AT, Popken DA, Kaplan AM, Plunkett LM, Becker RA, How well can in vitro data predict in vivo effects of chemicals? Rodent carcinogenicity as a case study, Regul. Toxicol. Pharmacol 77 (2016) 54–64. [PubMed: 26879462]

[17]. Sipes NS, Martin MT, Reif DM, Kleinstreuer NC, Judson RS, Singh AV, Chandler KJ, Dix DJ, Kavlock RJ, Knudsen TB, Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data, Toxicol. Sci 124 (1) (2011) 109–127. [PubMed: 21873373]

[18]. Watford S, Ly Pham L, Wignall J, Shin R, Martin MT, Friedman KP, ToxRefDB version 2.0: improved utility for predictive and retrospective toxicology analyses, Reprod. Toxicol 89 (2019) 145–158. [PubMed: 31340180]

[19]. Grenet I, Comet JP, Schorsch F, Ryan N, Wichard J, Rouquie D, Chemical in vitro bioactivity profiles are not informative about the long-term in vivo endocrine mediated toxicity, Comput. Toxicol 12 (2019).

[20]. Becker RA, Dreier DA, Manibusan MK, Tony Cox LA, Simon TW, Bus JS, How well can carcinogenicity be predicted by high throughput “characteristics of carcinogens” mechanistic data? Regul. Toxicol. Pharmacol (2017).

[21]. Honda GS, Pearce RG, Pham LL, Setzer RW, Wetmore BA, Sipes NS, Gilbert J, Franz B, Thomas RS, Wambaugh JF Using the concordance of in vitro and in vivo data to evaluate extrapolation assumptions. PLoS One. 2019;14(5): e0217564. adherence to PLOS ONE policies on sharing data and materials, per contractual agreements with the United States Environmental Protection Agency regarding data availability and transparency of publicly funded research. [PubMed: 31136631]

[22]. Paul Friedman K, Gagne M, Loo LH, Karamertzanis P, Netzeva T, Sobanski T, Franzosa JA, Richard AM, Lougee RR, Gissi A, Lee JJ, Angrish M, Dorne JL, Foster S, Raffaele K, Bahadori T, Gwinn MR, Lambert J, Whelan M, et al., Utility of In vitro bioactivity as a lower bound estimate of in vivo adverse effect levels and in risk-based prioritization, Toxicol. Sci 173 (1) (2020) 202–225. [PubMed: 31532525]

[23]. Klaren WD, Ring C, Harris MA, Thompson CM, Borghoff S, Sipes NS, Hsieh JH, Auerbach SS, Rager JE, Identifying attributes that influence in vitro-to-in vivo concordance by comparing in vitro Tox21 bioactivity versus in vivo drugmatrix transcriptomic responses across 130 chemicals, Toxicol Sci. 167 (1) (2019) 157–171. [PubMed: 30202884]

[24]. EPA US. ToxCast Owner’s Manual - Guidance for Exploring Data 2019 [cited 2019 April 1]. Available from: https://www.epa.gov/sites/production/files/2018-04/documents/toxcastownermanual4252018.pdf.

[25]. Judson R, Houck K, Martin M, Richard AM, Knudsen TB, Shah I, Little S, Wambaugh J, Setzer RW, Kothiya P, Phuong J, Filer D, Smith D, Reif D, Rotroff D, Kleinstreuer N, Sipes N, Xia M, Huang R, et al., Analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space, Toxicol. Sci 153 (2) (2016) 409. [PubMed: 27605417]

[26]. Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, Xia M, Huang R, Rotroff DM, Filer DL, Houck KA, Martin MT, Sipes N, Richard AM, Mansouri K, Setzer RW, Knudsen TB, Crofton KM, Thomas RS, Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor, Toxicol Sci. 148 (1) (2015) 137–154. [PubMed: 26272952]

[27]. EPA US. The ToxCast(TM) Analysis Pipeline(tcpl) An R Package for Processing and Modeling Chemical Screening Data (Version 2.0) 2019 [cited 2019 Jun 1]. Available from: https://cran.r-project.org/web/packages/tcpl/vignettes/Introduction_Appendices.html.

[28]. Armitage JM, Wania F, Arnot JA, Application of mass balance models and the chemical activity concept to facilitate the use of in vitro toxicity data for risk assessment, Environ. Sci. Technol 48 (16) (2014) 9770–9779. [PubMed: 25014875]

[29]. NCATS. TOX21 PUBLIC DATA 2020 [cited 2020 Oct 1]. Available from: https://tripod.nih.gov/tox21/assays/.

[30]. Corning. Surface areas and guide for recommended medium volumes for corning cell culture vessels 2020 [cited 2020 July 1]. Available from: https://www.corning.com/catalog/cls/documents/application-notes/CLS-AN-209.pdf.

[31]. NTP. DrugMatrix National Toxicology Program (NTP)2017 [cited 2017 5 April]. Available from: https://ntp.niehs.nih.gov/drugmatrix.

[32]. Baldi P, Long AD, A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes, Bioinformatics 17 (6) (2001) 509–519. [PubMed: 11395427]

[33]. Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, Yamada H. Open TG-GATEs: a large-scale toxicogenomics database. Nucleic Acids Res. 2015; 43(Database issue):D921–7. [PubMed: 25313160]

[34]. TG-Gates. Open TG-Gates Life Science Database Archive 2019 [cited 2019 Jan 15]. Available from: https://dbarchive.biosciencedbc.jp/en/open-tggates/download.html.

[35]. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. [PubMed: 25605792]

[36]. Institute B. MSigDB: Molecular Signatures Database 2019 [cited 2019 Oct 1]. Available from: https://www.gsea-msigdb.org/gsea/msigdb/index.jsp.

[37]. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U.S.A 102 (43) (2005) 15545–15550. [PubMed: 16199517]

[38]. Varemo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. 2013;41(8):4378–91. [PubMed: 23444143]

[39]. Farmahin R, Williams A, Kuo B, Chepelev NL, Thomas RS, Barton-Maclaren TS, Curran IH, Nong A, Wade MG, Yauk CL, Recommended approaches in the application of toxicogenomics to derive points of departure for chemical risk assessment, Arch. Toxicol 91 (5) (2017) 2045–2065. [PubMed: 27928627]

[40]. Rager MJE, Ring CL, Fry RC, Suh M, Proctor DM, Haws LC, Harris MA, Thompson CM, High-Throughput Screening Data Interpretation in the Context of In Vivo Transcriptomic Responses to Oral Cr(VI) Exposure, Toxicol. Sci (2017).

[41]. Mischler A, Karakis V, Mahinthakumar J, Carberry CK, San Miguel A, Rager JE, Fry RC, Rao BM, Two distinct trophectoderm lineage stem cells from human pluripotent stem cells, J. Biol. Chem 100386 (2021). [PubMed: 33556374]

[42]. Rager JE, Auerbach SS, Chappell GA, Martin E, Thompson CM, Fry RC, Benchmark dose modeling estimates of the concentrations of inorganic arsenic that induce changes to the neonatal transcriptome, proteome, and epigenome in a pregnancy cohort, Chem. Res. Toxicol 30 (10) (2017) 1911–1920. [PubMed: 28927277]

[43]. Rager JE, Yosim A, Fry RC Prenatal exposure to arsenic and cadmium impacts infectious disease-related genes within the glucocorticoid receptor signal transduction pathway. Int J Mol Sci. 2014;15(12):22374–91. [PubMed: 25479081]

[44]. Rager JE, Bailey KA, Smeester L, Miller SK, Parker JS, Laine JE, Drobna Z, Currier J, Douillet C, Olshan AF, Rubio-Andrade M, Styblo M, Garcia-Vargas G, Fry RC, Prenatal arsenic exposure and the epigenome: altered microRNAs associated with innate and adaptive immune signaling in newborn cord blood, Environ. Mol. Mutagen 55 (3) (2014) 196–208. [PubMed: 24327377]

[45]. Rager JE, Moeller BC, Miller SK, Kracko D, Doyle-Eisele M, Swenberg JA, Fry RC, Formaldehyde-associated changes in microRNAs: tissue and temporal specificity in the rat nose, white blood cells, and bone marrow, Toxicol Sci. 138 (1) (2014) 36–46. [PubMed: 24304932]

[46]. Rager JE, Moeller BC, Doyle-Eisele M, Kracko D, Swenberg JA, Fry RC, Formaldehyde and epigenetic alterations: microRNA changes in the nasal epithelium of nonhuman primates, Environ. Health Perspect 121 (3) (2013) 339–344. [PubMed: 23322811]

[47]. Pearce RG, Setzer RW, Strope CL, Sipes NS, Wambaugh JF, httk: R package for high-throughput toxicokinetics, J Stat Soft. 79 (4) (2017) 1–26.

[48]. EPA US. Chemistry Dashboard 2019 [cited 2019 Jan 15]. Available from: https://comptox.epa.gov/dashboard/.

[49]. Sipes NS, Wambaugh JF, Pearce R, Auerbach SS, Wetmore BA, Hsieh JH, Shapiro AJ, Svoboda D, DeVito MJ, Ferguson SS, An intuitive approach for predicting potential human health risk with the Tox21 10k library, Environ. Sci. Technol 51 (18) (2017) 10786–10796. [PubMed: 28809115]

[50]. Chawla NV, Bowyer KW, Hall LA, Kegelmeyer WP, SMOTE: synthetic minority oversampling technique, J. Artificial Intelligence Res 16 (2002) 321–357.

[51]. Breiman L, Random forests, Machine Learning. 45 (1) (2001) 5–32.

[52]. Liaw A, Wiener M, Classification and regression by randomForest, R News. 2 (3) (2002) 18–22.

[53]. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77. [PubMed: 21414208]

[54]. Welling SH, Refsgaard HF, Brockhoff PB, Clemmensen LH Forest Floor Visualizations of Random Forest. arXiv preprint arXiv:160509196. 2016:1–26.

[55]. Ring C, Sipes NS, Hsieh JH, Carberry C, Koval LE, Klaren WD, Harris MA, Auerbach SS, Rager JE, Dataset for Predictive Modeling of Biological Responses in the Rat Liver using In Vitro Tox21 Bioactivity: Benefits from High-Throughput Toxicokinetics, UNC Dataverse, Ragerlab-Dataverse. 2020. Available at: 10.15139/S3/WCLFWZ.

[56]. Palczewska A, Palczewki J, Robinson RM, Neagu D, Interpreting random forest classification models using a feature contribution method, Integr. Resuable Syst (2014) 193–218.

[57]. Stanley LA, Horsburgh BC, Ross J, Scheer N, Wolf CR, PXR and CAR: nuclear receptors which play a pivotal role in drug disposition and chemical toxicity, Drug Metab Rev. 38 (3) (2006) 515–597. [PubMed: 16877263]

[58]. Fischer FC, Henneberger L, Konig M, Bittermann K, Linden L, Goss KU, Escher BI, Modeling exposure in the Tox21 in vitro bioassays, Chem. Res. Toxicol 30 (5) (2017) 1197–1208. [PubMed: 28316234]

[59]. Gulden M, Morchel S, Seibert H, Factors influencing nominal effective concentrations of chemical compounds in vitro: cell concentration, Toxicol. In Vitro 15 (3) (2001) 233–243. [PubMed: 11377096]

[60]. Casey WM, Chang X, Allen DG, Ceger PC, Choksi NY, Hsieh JH, Wetmore BA, Ferguson SS, DeVito MJ, Sprankle CS, Kleinstreuer NC, Evaluation and optimization of pharmacokinetic models for in vitro to in vivo extrapolation of estrogenic activity for environmental chemicals, Environ. Health Perspect 126 (9) (2018) 97001. [PubMed: 30192161]

[61]. Kleinstreuer NC, Ceger P, Watt ED, Martin M, Houck K, Browne P, Thomas RS, Casey WM, Dix DJ, Allen D, Sakamuru S, Xia M, Huang R, Judson R, Development and validation of a computational model for androgen receptor activity, Chem. Res. Toxicol 30 (4) (2017) 946–964. [PubMed: 27933809]

[62]. Kosnik MB, Strickland JD, Marvel SW, Wallis DJ, Wallace K, Richard AM, Reif DM, Shafer TJ, Concentration-response evaluation of ToxCast compounds for multivariate activity patterns of neural network function, Arch. Toxicol 94 (2) (2020) 469–484. [PubMed: 31822930]

[63]. Zakharov AV, Peach ML, Sitzmann M, Nicklaus MC, QSAR modeling of imbalanced high-throughput screening data in PubChem, J. Chem. Inf. Model 54 (3) (2014) 705–712. [PubMed: 24524735]

[64]. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl 73 (2017) 220–239.

[65]. Pozzolo AD, Caelen O, Borgne YL, Waterschoot S, Bontempi G, Learned lessons in credit card fraud detection from a practitioner perspective, Expert Syst. Appl 41 (10) (2014) 4915–4928.

[66]. Franzosa JA, Bonzo JA, Jack J, Baker NC, Kothiya P, Witek RP, Hurban P, Siferd S, Hester S, Shah I, Ferguson SS, Houck KA, Wambaugh JF, High-throughput toxicogenomic screening of chemicals in the environment using metabolically competent hepatic cell cultures, NPJ Syst. Biol. Appl 7 (1) (2021) 7. [PubMed: 33504769]

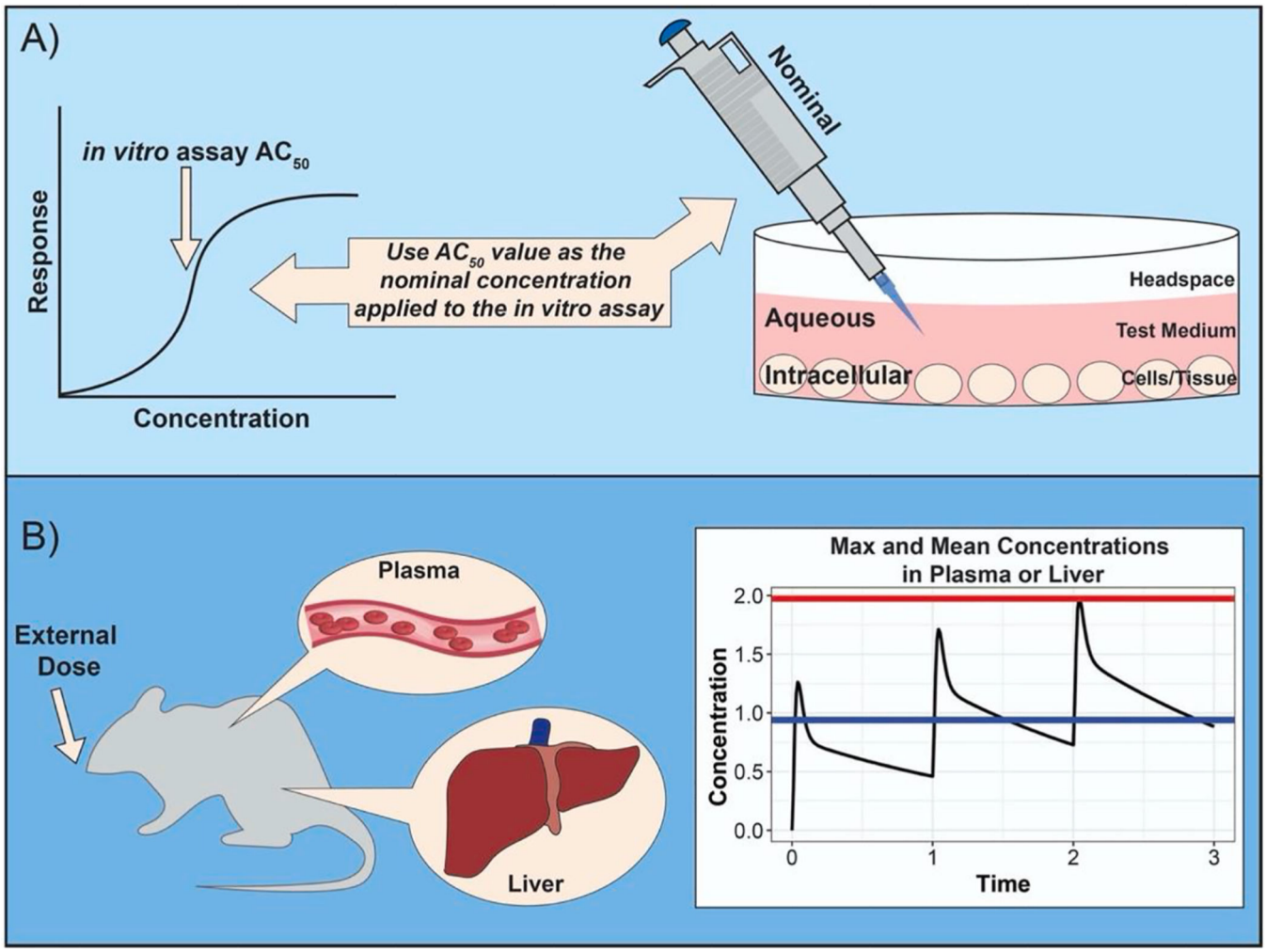

Fig. 1.

Concentration estimates tested as different predictor/outcome variables to identify those producing maximal predictivity. (A) *In vitro* chemical concentrations were evaluated as either the nominal concentrations (representing the concentration of the tested chemical dissolved in the solute applied to the assay), the aqueous concentrations (representing the concentration of the tested chemical that dissolves in the assay solution), or the intracellular concentrations (representing the concentration of the tested chemical that enters and accumulates in the cells). (B) *In vivo* chemical concentrations were evaluated as the concentration circulating within blood plasma or the concentration that is absorbed in the specific target tissue of interest (in this case study, the liver). Within circulating plasma or the specific target tissue, chemical concentration were estimated as the maximum or the mean concentration after exposure.

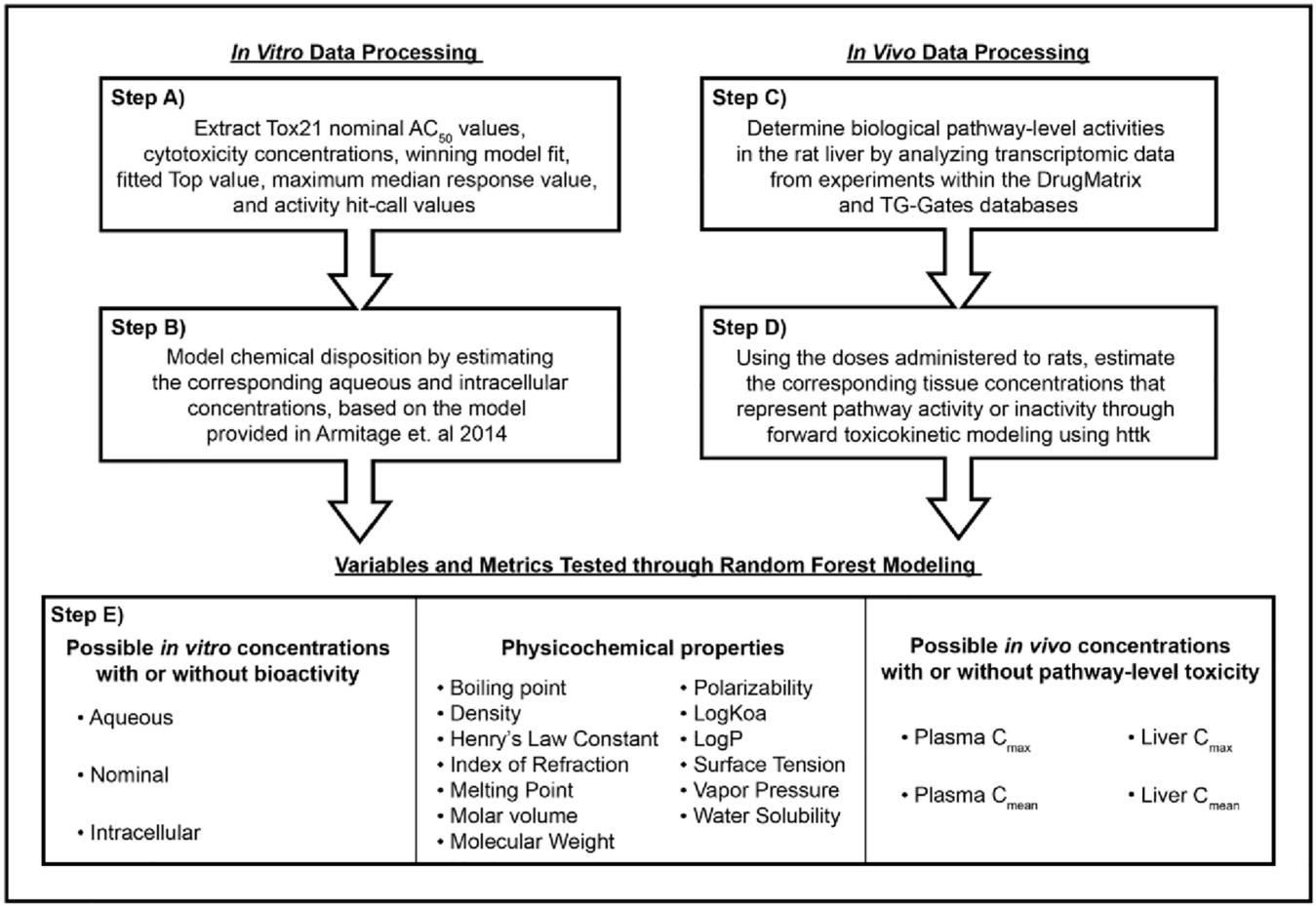

Fig. 2.

Study overview. Flowchart of steps carried out to evaluate predictivity of *in vitro* Tox21 bioactivity vs. *in vivo* pathway-level activities, through the incorporation of chemical disposition and toxicokinetic modeling.

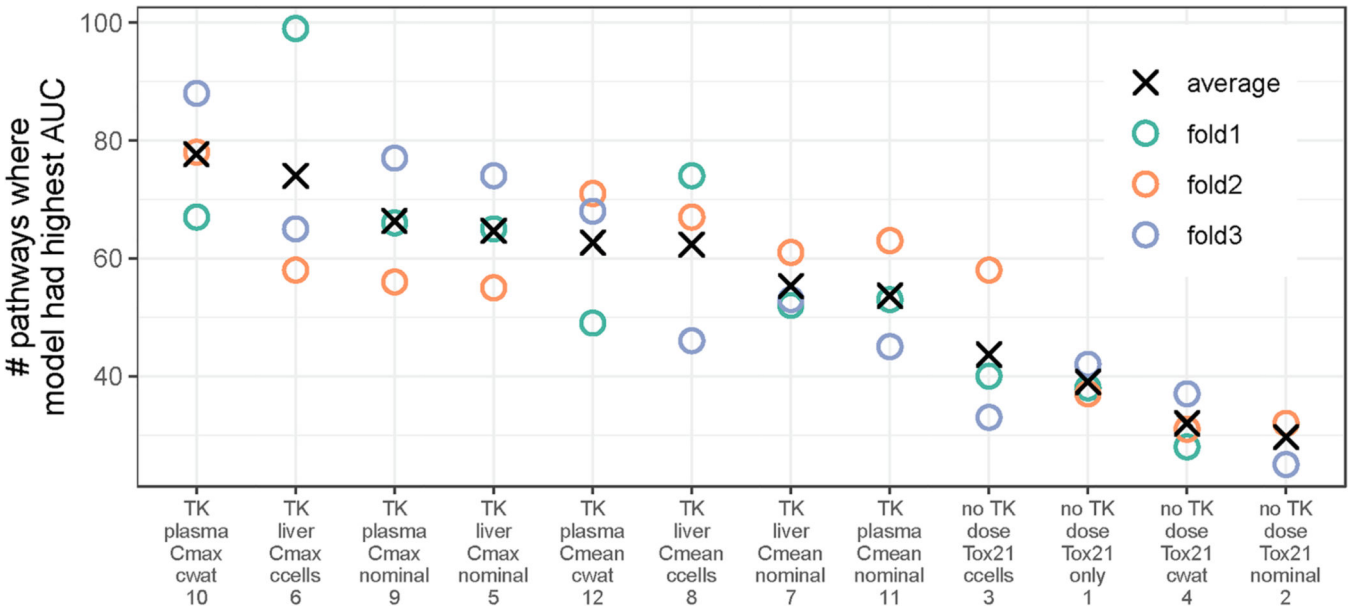

Fig. 3.

Model performance rankings for each combination of variables tested. Variable combinations were ranked based on the number of biological pathways where the corresponding model had the highest AUC-ROC. The models (variable combinations) on the horizontal axis are listed in order of descending number of pathways for which the corresponding model had the highest AUC-ROC, averaged across folds. Colored circles indicate the number of pathways for each fold; black X's indicates the average number of pathways across folds. Model numbers are provided, as summarized in Table 1. Each model is labeled as "TK" (including toxicokinetics) or "no TK" (not including toxicokinetics). Abbreviations: ccells (intracellular concentration); Cmean (mean tissue concentration); Cmax (maximum tissue concentration); cwat (aqueous [or water] concentration).

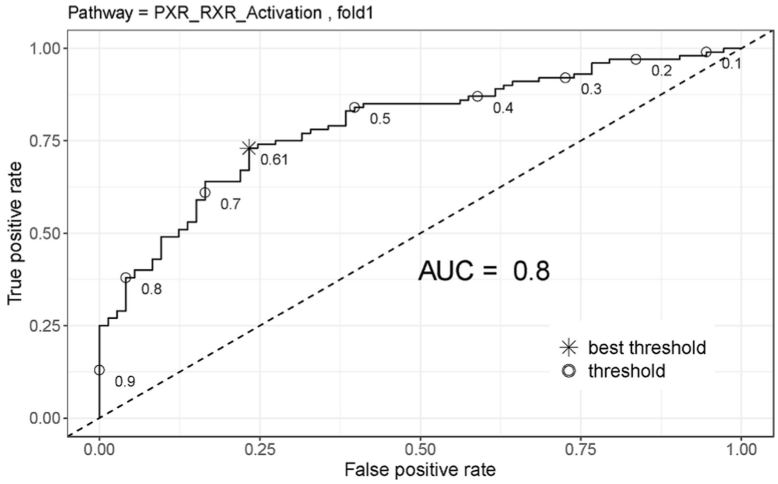

Fig. 4.

Example model performance metrics for the PXR-RXR activation pathway. An example ROC for the winning model for the fold-1 test set is displayed. AUC-ROC for this curve is 0.80. Points on the curve corresponding to possible classification thresholds between 0.1 and 0.9 are marked with open circles and the threshold value. The best threshold (0.61) is marked with a star; this is the optimal threshold to set for classifying an experiment as active (model-predicted activity probability above 0.61) or inactive (model-predicted activity probability below 0.61), maximizing the true positive rate and minimizing the false positive rate. The dashed diagonal line indicates the theoretical ROC for a perfectly useless classifier (AUC = 0.5). The ROC and the best threshold varied according to pathway and test set, with all values provided in supplementary material.

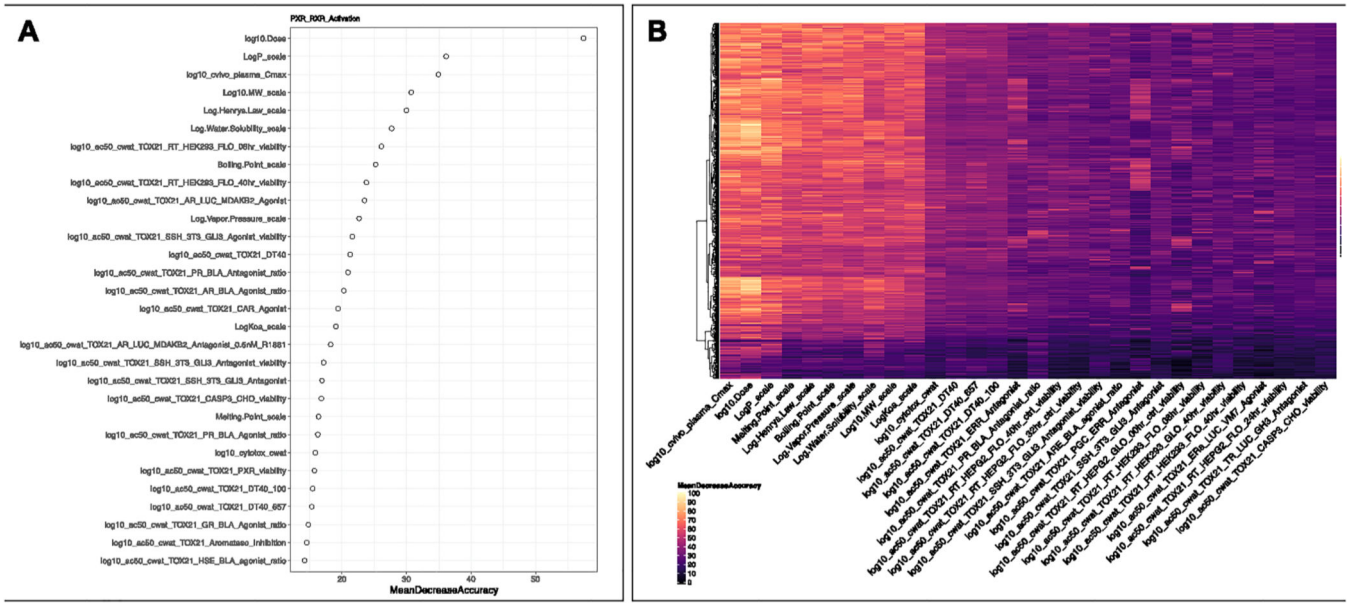

Fig. 5.

Predictor-variable importance for the winning random forest model. Results are summarized (A) for the example PXR-RXR activation pathway, with variables are arranged from top to bottom in decreasing order of mean decrease in accuracy in predicting out-of-bag samples when each predictor variable was permuted; mean decrease in accuracy values are indicated along the x-axis. Results are also summarized (B) across all evaluated pathways, where a heatmap shows mean decrease in accuracy in predicting out-of-bag samples for each pathway (rows) when each variable (columns) is perturbed. Variables are arranged from left to right in decreasing order of average mean decrease in accuracy across pathways. These plots focus on the top 30 variables.

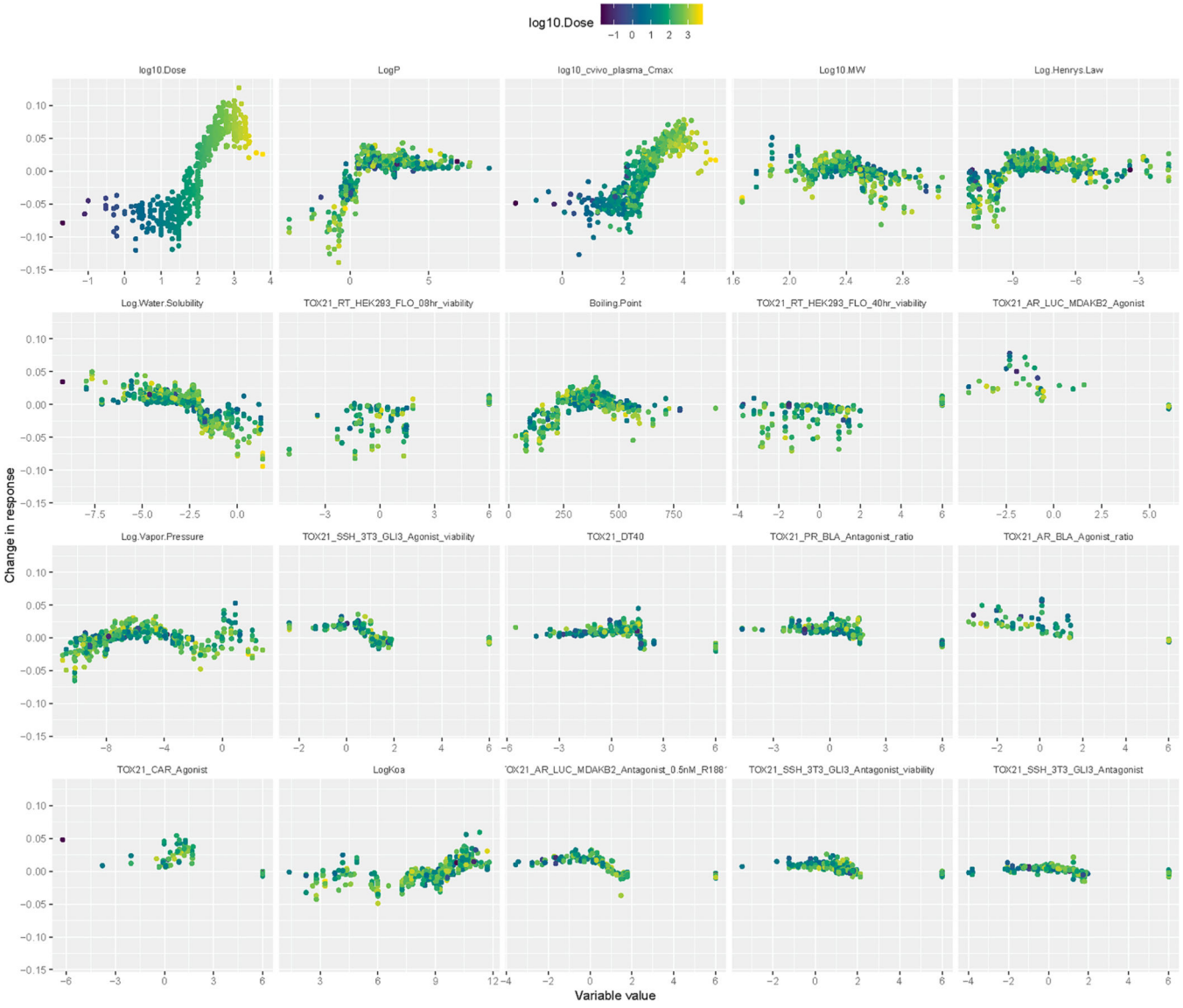

Fig. 6.

Example feature contribution plots. Results are shown for the top 20 most important predictor variables (as shown in Fig. 5) for the winning random forest model for the PXR-RXR activation pathway. Each panel shows the contribution of the specified predictor variable to the probability that each experiment is classified as active in this pathway; each point represents one *in vivo* experiment. Each panel includes the same number of points (*in vivo* experiments); however, points may be plotted on top of one another, so that fewer individual points may be visible in some panels. For *in vitro* Tox21 $\log_{10}$ AC_{50} values (variables beginning with "TOX21"), placeholder values of 99 (for hit-call 0) are plotted at the value 6. Points are color-coded by a specific predictor variable of interest; namely, $\log_{10}$ administered dose for each experiment. This coloration allows for the qualitative assessment of whether each predictor variable may have an interaction with the response variable, $\log_{10}$ dose, indicating the concentration estimated to elicit activation of the PXR-RXR pathway in the rat liver.

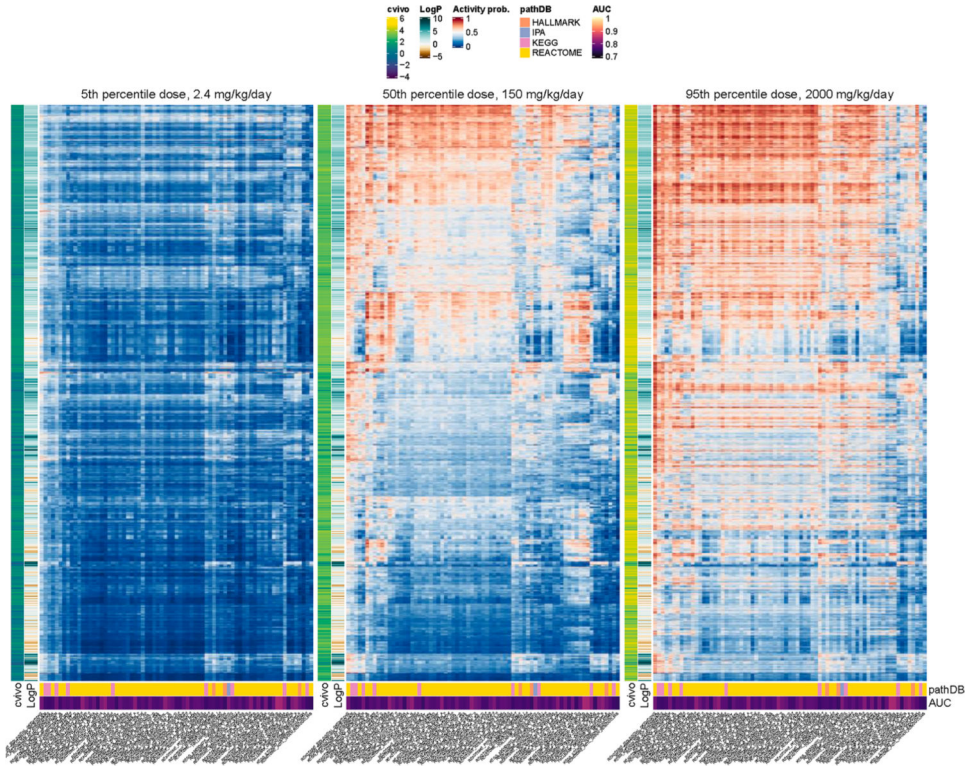

Fig. 7.

In vivo pathway activity predictions across Tox21 chemicals. Results are shown for pathways with the highest model performance. These heat maps display the probability of each pathway being altered *in vivo* after exposure to a chemical in Tox21, based on the winning random forest model. Rows represent 6617 Tox21 chemicals; columns represent the 73 pathways with the top 10% AUC-ROC (pathway names are truncated to 30 characters). Rows (chemicals) are annotated with colorbars representing $\log_{10}$ of maximum plasma concentration ("Cvivo") and log P. Columns (pathways) are annotated with colorbars representing pathway database ("pathDB") and AUC-ROC ("AUC"). Pathway activation predictions were generated based on treatment doses spanning the (left) 5th, (middle) 50th, and (right) 95th percentiles of the doses tested in DrugMatrix/TG-Gates. For visualization purposes, only the pathways demonstrating the highest predictive performance (i.e., top 10% AUC-ROC) are included. For all predictions, see supplementary material (S12 File, S13 File, S14 File)

Table 1

Combinations of predictor variables that were included in the 12 total random forest classifier models. These combinations of predictor variables were designed to evaluate the potential influence of toxicokinetics (TK) on the overall predictivity of Tox21 bioactivity towards predicting *in vivo* pathway-level activities in the rat liver. Note that randomly permuted versions of each model were also developed to evaluate model performance, yielding a total of 24 models.

Predictor Variables	**TK Excluded**				**TK Included**							
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**	**Model 10**	**Model 11**	**Model 12**
Physicochemical Properties	X	X	X	X	X	X	X	X	X	X	X	X
Log_{10} Dose	X	X	X	X	X	X	X	X	X	X	X	X
***In Vitro* Tox21 Data**												
Binary Hit-Call	X											
Nominal log_{10} AC_{50}		X			X		X		X		X	
Intracellular log_{10} AC_{50}			X			X		X				
Aqueous log_{10} AC_{50}				X						X		X
Nominal log_{10} Cytotoxicity Point		X			X		X		X		X	
Intracellular log_{10} Cytotoxicity Point			X			X		X				
Aqueous log_{10} Cytotoxicity Point				X						X		X
***In Vivo* DrugMatrix or TG-Gates Concentration Values (C_{vivo})** *												
TK-Predicted Liver Max					X	X						
TK-Predicted Liver Mean							X	X				
TK-Predicted Plasma Max									X	X		
TK-Predicted Plasma Mean											X	X

*C_{vivo} are the TK-predicted tissue concentrations corresponding to the daily dose, dosing duration, and dosing route reported for each DrugMatrix or TG-Gates experiment.