ORD MANUSCRIPT COVER SHEET

**Title:** Collaborative Evaluation of *In silico* Predictions for High Throughput Toxicokinetics

**Authors:**

|  |  |  |
| --- | --- | --- |
| John Wambaugh | 1 EPA/ORD/CCTE | [Wambaugh.john@epa.gov](mailto:Wambaugh.john@epa.gov) |
| Nisha Sipes | 1 | [Sipes.nisha@epa.gov](mailto:Sipes.nisha@epa.gov) |
| Kamel Mansouri | 2 Innotiv  3 NICEATM | kamel.mansouri@nih.gov |
| Jon Arnot | 4 Arnot Research & Consulting | jon@arnotresearch.com |
| Trevor Brown | 4 | trevor.n.brown@gmail.com |
| Chris Cook | 1 | cookchristopher732@gmail.com |
| Daniel Dawson | 1 | ddawson@integral-corp.com |
| Sarah Davidson Fritz | 1 | DavidsonFritz.Sarah@epa.gov |
| John DiBella | 5 Simulations Plus | john.dibella@simulations-plus.com |
| Stephen Ferguson | 6 NIEHS | stephen.ferguson@nih.gov |
| Rocky Goldsmith | 1 | rgoldsmith@congruencetx.com |
| Chris Grulke | 1 | [cmgrulke@gmail.com](mailto:cmgrulke@gmail.com) |
| Richard Judson | 1 | Judson.Richard@epa.gov |
| Michael Lawless | 5 | mlawless@simulations-plus.com |
| Gilberto Padilla Mercado | 1 | PadillaMercado.Gilberto@epa.gov |
| Grace Patlewicz | 1 | Patlewicz.Grace@epa.gov |
| Ester Papa | 7 University of Insubria, Varese | ester.papa@uninsubria.it |
| Prachi Pradeep | 8 German Federal Institute for Risk Assessment (BfR) | prachipradeep.15@gmail.com |
| Alessandro Sangion | 2 | alessandro.sangion@mail.utoronto.ca |
| Risa Sayre | 1 | sayre.risa@epa.gov |
| Russell Thomas | 1 | Thomas.Russell@epa.gov |
| Rogelio Tornero-Velez | 1 | tornero-velez.rogelio@epa.gov |
| Barbara Wetmore | 1 | wetmore.barbara@epa.gov |
| Michael Devito | 1 | Devito.Michael@epa.gov |

**Internal Peer Reviewers:** maybe Todd Martin, Dan Chang, or Marina Evans?

**Target Journal:** Not sure, possibly: Computational Toxicology, Journal of Pharmacokinetics and Pharmacodynamics, Drug Metabolism and Disposition, Drug Discovery Today, Environmental Science & Technology (probably too many figures), maybe the special issue from QSAR2021 if Grace can get us in late

**Four bullet point summary:**

* Toxicokinetic (TK) information, such as elimination half-life (thalf, plotted below), is critical for understanding chemical risk
* Here we collected *in silico* (quantitative structure-property relationship, QSPR) predictions of key *in vitro* determinants of TK from several different models
* The models were evaluated for ability to reproduce *in vitro* and *In vivo* measurements of TK
* Overall, high throughput physiologically-based TK (PBTK) model performed similarly when using TK QSPRs as when the actual *in vitro* measured data were used

**One sentence description:** This collaborative trial demonstrates that multiple QSPRs exist that make reasonably accurate chemical structure-based predictions for *in vitro* TK parameters

**Chemicals Involved:** 87 ToxCast chemicals with diverse uses for which high-throughput toxicokinetic (HTTK) data are available

**Collaborative Evaluation of *In Silico* Predictions for High Throughput Toxicokinetics**

John Wambaugh1,\*, Nisha Sipes1, Jon Arnot2, Trevor Brown2, Christopher Cook1,3, Daniel Dawson1, Sarah Davidson1, , John DiBella4, Stephen Ferguson5, Rocky Goldsmith1, Chris Grulke1, Richard Judson1, Michael Lawless4, Kamel Mansouri6, Grace Patlewicz1, Ester Papa7, Prachi Pradeep1,8, Alessandro Sangion2, Risa Sayre1, Russell Thomas1, Rogelio Tornero-Velez1, Barbara Wetmore1, and Michael Devito1

1. Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 28311
2. ARC Arnot Research and Consulting Inc.
3. Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee 38331
4. Simulations Plus, Inc.,
5. Division of the National Toxicology Program (NTP), National Institute of Environmental Health Sciences, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 28317
6. NTP Interagency Center for the Evaluation of Alternative Toxicological Methods, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 28317
7. Department of Theoretical and Applied Sciences, University of Insubria, Varese
8. German Federal Institute for Risk Assessment (BfR)

\*Corresponding Author: wambaugh.john@epa.gov

# Abstract

To assess public health risks posed by chemicals we need to understand chemical absorption, distribution, metabolism, and elimination by the body (that is, toxicokinetics or TK). Unfortunately, chemical-specific TK data are often unavailable*. In silico* predictions along with high throughput TK (HTTK) methods have the potential to address this gap. This collaborative trial uses a database of *in vivo* measured toxicokinetic data to evaluate *in silico* approaches for HTTK. Six different sets of quantitative structure-property-relationship (QSPR) tools for predicting TK were evaluated. QSPR predictions were evaluated by using the predicted parameter values within a high throughput physiologically based TK (PBTK) model to predict *in vivo* measured plasma concentrations for 92 chemicals, mostly in rats. Root mean squared log10 error (RMSLE) was calculated for toxicologically-relevant dosimetry statistics AUC (time-integrated area under the curve) and Cmax (peak concentration. For Cmax optimal fits to the *in vivo* data indicated a best case RMSLE of 0.5. PBTK using chemical-specific *in vitro* data predicted AUC with RMSLE 0.74, while QSPRs ranged in accuracy from 0.74 to 0.84. For AUC *in vivo* parameters gave RMSLE 0.5. Chemical-specific *in vitro* data had RMSLE 1.3 and QSPRs ranged from 1.3-1.4. *In vivo* parameters for the full concentration time-course data had RMSLE of 0.62. Predictions based on *in vitro* HTTK had RMSLE of 1.18. RMSLE for individual QSPRs ranged from 1.19 to 1.3. QSPRs predict plasma binding well (RMSLE 0.03 – 0.07) but have difficulty predicting metabolic clearance (RMSLE 0.37 – 1.28). However, a consensus prediction using the maximum clearance predicted across all QSPRs predicted AUC with RMSLE 1.03 and full time-course with RMSLE 1.09. The consensus predictions outperformed the *in vitro* measured data for the evaluation chemicals. For novel compounds a consensus QSPR approach may yield a reasonable prediction. This evaluation characterizes the accuracy of HTTK approaches for new chemicals based on both *in vitro* measurement and structure-based *in silico* predictions.

# Introduction

Toxicokinetics (TK) describes the absorption, distribution, metabolism, and excretion (ADME) of a chemical compound in the body as a function of time [1]. Since TK allows the prediction of internal tissue concentrations as a function of chemical exposure it provides critical information for assessing risk posed by a chemical to public health [2]. TK allows interpretation of biomonitoring data [3], dosimetric anchoring of animal toxicity studies [4], and quantitative *in vitro*-*in vivo* extrapolation (or IVIVE) from high throughput bioactivity studies [5]. These *in vitro* bioactivity data are available for thousands of chemicals (for example, the ToxCast [6] and Tox21 [7] screening programs). IVIVE of bioactivity data relies upon TK to relate the concentrations found to be active *in vitro* with doses that could cause these concentrations in tissues [8-12]. Unfortunately, chemical-specific information on TK are often unavailable for thousands of chemicals in commerce and the environment [13]. New approach methodologies (NAMs) are being developed throughput the chemical risk assessment process [14] including new methods for assessing TK [15,16].

For non-therapeutic compounds TK data were traditionally developed using animal studies; for reasons of both ethics and resources these studies are no longer desirable nor practical for the thousands of remaining chemicals [17]. An alternative technology developed by the pharmaceutical industry relies on characterizing certain aspects of TK *in vitro* and then extrapolating to *in vivo* conditions to estimate TK parameters such as AUC, often within a three-fold error [18]. For the past decade, government chemical regulatory agencies working with their collaborators and contractors have been collecting chemical-specific *in vitro* data that allow prediction of TK [8-12,19]. These data currently span approximately one thousand chemicals used for industry, pesticides, diet, therapy, and consumer products. However, thousands more remain uncharacterized.

The U.S. National Academies of Science, Engineering, and Mathematics have recognized that *in vitro* TK data “enabled first-tier risk-based rankings of chemicals on the basis of margins of exposure—the ratio of exposures that cause effects (or bioactivity) to measured or estimated human exposures” [20]. To address the remaining chemicals for which even *in vitro* TK data are unavailable, multiple organizations have developed *in silico* quantitative structure-property relationship (QSPR) models for predicting these values [21-25]. Meanwhile, EPA and other organizations are continuing to accumulate chemical-specific TK data, both *in vivo* – that is, curation of data from the scientific literature as well as a targeted animal studies only when needed [26,27] – as well as *in vitro* [11,19,12].

Here we examine the TK predictions from six different modeling teams. Four modeling teams produced QSPR models for two key toxicokinetic parameters that can be measured *in vitro*: intrinsic hepatic clearance (Clint measured with hepatocyte incubations, [28]) and fraction unbound in plasma (fup, typically measured via rapid equilibrium dialysis [29]). The models are initially evaluated for their ability to reproduce *in vitro* measured values. However, we focus on analysis of the predictions for chemical concentration as a function of time (CvT) that can be made when a generic physiologically-based TK (PBTK) model [30] is used with the predictions from each QSPR. For this study *in vivo* plasma and blood concentration vs. time data for rat and human were initially available for 101 chemicals from the CvTdb [27]. Models are evaluated for ability to reproduce the full CvT curve as well as summary statistics (such as peak plasma concentration) and parameters (such as chemical half-life). Two additional models for chemical half-life are also evaluated. The six modeling teams were provided with chemical identities and some physico-chemical descriptors but were not provided with the actual *in vivo* evaluation data.

# Methods

As summarized in Table 1, three levels of evaluation were made. First (Level 1), QSPR predictions of parameters were compared against *in vitro*-measured values for each chemical where data were available. Then (Level 2) the ability of a PBTK model to predict chemical concentration vs. time behavior was evaluated against *in vivo* measurements for the full time-course observed for each combination of QSPR and chemical. Finally (Level 3) the ability of the QSPRs and PBTK model to predict summary statistics (for example Cmax, AUC, half-life) was evaluated. All analyses were performed in the free, open-source statistical analysis language R [31] v4.4.0. Analyses were performed on a Dell Precision 7560 laptop personal computer. Scripts to perform all analyses are available in RMarkdown [32] format as supplemental material.

## HTTK

R package “httk” [30] v2.3.2 was used for this analysis. “httk” can parameterize a physiologically based toxicokinetic (PBTK) model based on chemical-specific values for fraction unbound in plasma (fup, unitless) and intrinsic hepatic clearance (Clint, µL/min/106 hepatocytes). These experimentally measured values are collected from peer-reviewed literature and provided by “httk”. All the QSPRs analyzed here were trained to human data.

The generic PBTK model used here (model “pbtk”) consists of well-mixed compartments for the gut, kidney, liver, and rest of body. The model is parameterized for a chemical using fup and Clint plus equilibrium tissue:plasma partition coefficients predicted with a modified Schmitt’s method [33,34]. The model simulates both oral and intravenous dosing. Oral dosing is subject to first-pass metabolism by the liver before the compound distributes systemically. Among other species, the model includes physiological information for parameterizing both humans and rats (primarily from [35,33,36]. The model was simulated using command “httk::solve\_pbtk()” with option default.to.human=TRUE – that is, since no rat-specific values are predicted by the models under evaluation, comparisons to data from rats were done using rat physiology but human in vitro TK parameters.

The steady-state ratio of the concentration of chemical in blood and plasma (Rb:p) is an important parameter in the PBTK model and is used for converting between plasma predictions and observations in whole blood. The function httk::get\_rblood2plasma() either retrieves measured values of Rb:p from the literature or predicts the ratio by predicting the red blood cell:plasma equilibrium partition coefficient and then using the hematocrit fraction for the relevant species.

R is an interpreted language (primarily operated by a user from the command-line although scripts are common). The user can alter the values within a table which stores the fup and Clint values for all chemicals (httk::chem.phys\_and\_invitro.data). After alteration, the httk function will proceed using the new values in the table. The HTTK data can be returned to their default values via the command “httk::reset\_httk()”. By default, no QSPR values are included in the table. However, predictions can be loaded with the commands “httk::load\_sipes2017(overwrite=TRUE)”, “httk::load\_pradeep2020(overwrite=TRUE)”, or “httk::load\_dawson2021(overwrite=TRUE)” [23-25]. The argument “overwrite=TRUE” is needed so that *in vitro* measured data are overwritten whenever a chemical-specific prediction is available. To facilitate comparisons, a custom function “clear\_httk()” is included in the supplemental material which deletes all Human Clint and fup values. The OPERA predictions are available as Supplemental Table 1.

## QSPR Models

The QSPR models evaluated are summarized in Table 2. Four different modeling teams previously [23-25] produced quantitative structure-property relationship (QSPR) models for two key toxicokinetic parameters that can be measured *in vitro*: intrinsic hepatic clearance (Clint measured with hepatocyte incubations) and fraction unbound in plasma (fup). The consensus QSPR was constructed for fup by taking the inverse logit of the mean of the logit-transformed fup from each QSPR. For Clint the maximum value predicted by any QSPR was used. Two additional models for chemical half-life were also evaluated. Individual model predictions are available in Supplemental Table 1.

Due to the potential presence of the model evaluation data (that is, measured *in vitro* parameters) in the training sets for some or all the models, we attempted to remove predictions that seemed more like a direct retrieval of the chemical-specific values from a training set. Model predictions were removed for a particular model-chemical combination if both the predictions for Clint and fup were within 1% absolute fold error of the measured values, the predictions were omitted from the evaluations. Omitted predictions are listed in Supplemental Table 6. When a QSPR model prediction was missing for a particular chemical the mean prediction of the other models was used for evaluation purposes.

QSAR training set summary data

* number chemicals, max, min, average values
* will use the averages for missing values

## *In vitro* Data

For 61 of the study chemicals *in vitro* measurements were also available for comparison. These data are collected by the R package “httk” but are drawn from the peer reviewed scientific literature (including [9,5,12]). It is likely that these data were in the training set of some of the QSPR models. The *in vitro* measured values are available in Supplemental Table 1.

## *In vivo* Data

EPA has developed a public database of concentration vs. time data for building, calibrating, and evaluating TK models [27]. Curation and development of the database are ongoing, but when this study began there were 101 chemicals with either rat or human *in vivo* blood or plasma concentration vs. time data. The *in vivo* measured concentration vs. time values are available as Supplemental Table 2.

## Compartmental Model Fits

For chemicals with CvT data, parameters were estimated for empirical one- and two-compartment toxicokinetic models using R package “invivoPKfit” (<https://github.com/USEPA/CompTox-ExpoCast-invivoPKfit>). Three models were considered: one- and two-compartment empirical TK models and a flat “null hypothesis” where there was no systematic change in concentration vs. time. The model with the lowest Akaike Information Criterion (AIC) value – indicating model parsimony – was selected [37]. Data sets where the flat model was selected were omitted from further analysis. The empirical model fit was then used as a “best case” prediction scenario for comparison with PBTK parameterize by *in vitro* or QSPR predictions.

For both models a half-life was calculated from the terminal elimination rate as thalf = ln(2)/kelim. For the two-compartment model the volume of distribution at steady-state was used as Vd. For both models, clearance was calculated as Cltot = Vd \* kelim. The estimated TK parameters for both models are provided as Supplemental Table 3.

## Evaluation Metrics

Multiple statistics were used to evaluate predictions (*pred*) relative to observed values (*obs*) as appropriate. Relative Predictive Error (RPE) was calculated as RPE =, where if the observed value was 0 then the error was set to zero. We note that if the predicted value is 0 then RPE = -1. Absolute Average Fold Error (AAFE) was calculated as AAFE = , where if pred=0 and obs=0 we assigned = 0. Root Mean Squared Log Error (RMSLE) was calculated as RMSLE = .

Kolmogorov-Smirnov tests were performed using R function *ks.test*.

# Results

## Evaluation Chemicals and Predictions

There are 101 chemicals present in the public release of CvTdb (Sayre, 2020) that had plasma concentration data following either oral or intravenous doses given to rats or humans. These chemicals included: 57 from the Toxic Substances Control Act (TSCA) active inventory [38], 20 pharmaceuticals, 24 pesticides, 99 that are found in consumer products, 7 per- and poly-fluorinated substances (PFAS) [39], and 64 that are part of the ToxCast screening program. Note that a chemical could be in more than one of these categories.

Data sets were modeled for their suitability to evaluate HTTK predictions by systematically optimizing parameters for one- and two-compartment models {Padilla, 2024}. Of the 101 chemicals, 3 were best described by the “flat” model indicating that the data sets were too noisy to estimate empirical TK parameters. Given that these chemicals were poorly described by basic TK models – potentially indicating problems with chemical analysis sensitivity -- these chemicals were withheld from subsequent analysis. The data for each remaining chemical could be described using either a one- or two-compartment empirical pharmacokinetic models. There were 12 chemicals where we only had oral data, so that we could not estimate the parameters needed to make novel predictions but were well-described by an empirical TK model. Separate parameter estimates were made for each combination of compound and species for which there were data. For each remaining chemical the better of the one or two compartment models was used on the basis of model parsimony.

## QSPR Predictions

The number of chemicals for which predicitons could be made (that is, domain of applicability) varied from QSPR model to QSPR model. Throughout this effort we have reported statistics on different subsets of chemicals. There was only a subset of chemicals with predictions from representing an intersection of the models.

Additionally, HTTK in vitro data were not available for all the chemicals with CVT data, so evaluations aimed at characterizing the predictivity of in vitro HTTK only relied on 63 chemicals. Second Finally, statistics are reported for the maximal number of chemicals for which the QSPR could make predictions. The subset of chemicals with overlap between all models is more rigorous statistically, however the superset of all chemicals with predictions is more rigorous chemically.

For each QSPR we removed predictions where the predicted values for a given chemical were within 1% for both fup and Clint assuming these values reflected the chemical data present in the training set and the model method allowing for recall of the measurements. This affected 21 chemicals as predicted by OPERA (see Supplement Materials Table SupTable-PossibleTrainingChems.txt).

We summarize the chemical-specific properties and predictions in Figure 1. In Figure 1 similar chemicals (rows) and properties/predictions (columns) are clustered together based on Euclidean distance. All properties/predictions were centered (mean changed to zero) and scaled (divided by standard deviation) such that the value reflects the number of standard deviations from the mean. Interestingly, the first division between clusters in Figure 1 places all the Clint measurements and predictions on the one side and all the fup measurements and predictions on the other. The physico-chemical properties are divided between those two clusters, with Octanol:Water (partition coefficient, PC), Octanol:Air PC, Molecular Weight, Boiling Point, and Melting Point all clustering with Clint. Water solubility, vapor pressure, and the Henry’s law constant all clustered with fup.

## Level 1 Analysis

Our first level of evaluation directly compared the predictions of QSPR's with the *in vitro* measured values. Supplemental Table SupTable-QSPRPredCounts.txt gives the per QSPR number of chemicals that were made – the values ranged from a low of 55 for both fup and Clint for the Pradeep et al. (2020) QSPR to 70 for Clint predictions for both OPERA and IVBP. Evaluation data were available for Clint for 67 chemicals and fup for 62. Table 3 summarizes the fold errors by parameter and QSPR. Note that there were several potential training chemicals that were removed from OPERA because its nearest-neighbor algorithm predicted the values too accurately.

We evaluate model performance for Clint in Figure 2. Note that the QSPRS models perform reasonably similarly despite that the Dawson (2021) model is categorical (that is, predicting only three values: very slow, slow, and fast) while the other models are continuous. Across the chemicals the RMSLE (root mean squared log10 error) for Clint predictions for the three QSPRs trained in part on chemicals from the ToxCast project (OPERA, Dawson (2021), Pradeep (2020)) were lower – 0.63, 0.73, and 0.75 respectively. For the QSPRs trained more broadly Simulations Plus ADMET predictor and IVBP had larger RSMLEs – 1.1 and 1.5. In particular, IVBP tends to overestimate the *in vitro* measured clearance when it predicts that there is clearance. In Table 3 we summarize the fold errors for the five QSPRs. QSPRs not trained to ToxCast chemicals tended to overestimate the clearance measured in vitro. Also shown in Table 3 is that the predictions of Clint range from 100x lower than experimental values (log10 folder error of -2) to 1000x higher (log10 fold error of 3). Each QSPR’s predictions are compared separately to observations in Supplemental Figure 1.

For fup, across the limited evaluation chemicals, the Pradeep (2020) model was most accurate with RMSLE 0.28, followed more distantly by Dawson (2021), Simulations Plus, and OPERA – 0.92, 1.0, and 1.4 respectively. Also shown in Table 3 is that for fup the predictions range from 30x too low to more than a million times overestimated (log10 fold error of 6).

We examined the distributions of fold errors between the predictions and the measured data using a Kolmogorov-Smirnov test. For both Clint and fup the only QSPRs that differed from the others were OPERA and IVBP. Those two QSPRs which had a significant (p-value < 0.05) difference between the distribution of predicted values and the distributions of the three other QSPRs as well as each other.

## Level 2 Analysis

We show all CvT curve fits and predictions on a per chemical, species, and route basis in the Supplemental materials (Supplemental Figures SupFig-ChembyQSPRCvTPlots.pdf).

All predicted values were used with a HTTK PBTK model to make predictions. We bracket QSPR model performance in three ways: First, we use the HTTK PBTK model with the actual *in vitro* measured values. In subsequent figures this is labeled as “HTTK-InVitro”. Next, for a best-case performance, we use empirical (one or two compartment toxicokinetic model) fits to the *in vivo* data, labelled as “In Vivo Fits”. The one and two compartment models are simpler than the high throughput PBTK model used for all other scenarios, but because they have been optimized to the *in vivo* evaluation data itself, they are expected to outperform the other approaches here. Finally, for worst case performance we use y-randomization so that the measured values across all chemicals in the R package “httk” library are scrambled and assigned to the incorrect chemicals, labelled “HTTK-YRandom”.

Figure 4 shows the full predicted time-courses for each set of model predictions as well as the actual *in vitro* data and empirical model fits. Across the QSPRs the predictions tended to be within a factor of ten (indicated by the dashed lines) with a bias toward over-predicting at low concentrations. We see chemicals where there are vertical bars in Fig X, indicating that the predicted concentrations were relatively constant (and low) over time while the observed concentrations changed. Typically these are chemicals where the CvT timecourse was especially biphasic, with an initial rapid decline and then long tails where low levels of the chemical remained. In the tails the models tend to underestimate concentration, even for the in vivo fits. The worst four chemicals for the in vivo fits were all cases where systemic oral bioavailanbiltiy was very low: 2-(Perfluorooctyl)ethanol, 2-Hydroxy-4-methoxybenzophenone, Bensulide, and Propyzamide.

Prediction error as characterized RMSLE is calculated first on a per chemical and method basis (error for predictions for a single method and a single chemical aggregating over all doses, routes, and time points). For each prediction method the mean RMSLE across all chemicals is. Generally error was consistent despite the subset of chemicals used (that is, 1) those chemicals with in vitro HTTK data, 2) those without in vitro data, 3) those chemicals representing the intersection of all QSPR models, and 4) all chemicals predicted by a given method. As summarized in Table 4, optimal performance is given by the empirical fits to the data, with a RMSLE between 0.55 and 0.69. Worst case performance is given by the y-randomized measured data with a RMSLE between 1.2 and 1.4. Performance of HTTK using the in vitro measured parameters is closer to the y-randomization than the in vivo fits – RMSLE of 1.2 indicates slightly more than a factor of ten on average.

The QSPRs perform roughly equally, with RMSLE generally ranging between 1.1 and 1.3 in Table 4. The OPERA model performs exceptionally well for the chemicals without in vitro data, with a RMSLE of 093. Notably the QSPRs are close to the performance of the in vitro HTTK data. Remarkably, the consensus predictor (using the mean plasma binding and the maximal predicted metabolism) outperforms the in vitro HTTK data. Absolute Average Fold Error results are summarized in Supplemental Table SupTable-Level2-Cvt-AAFEstats.txt.

In Figure 5 we examine the distribution of per chemical RMSLE. In the first panel of Figure 5 all observed time points are valued equally, without consideration of phase (absorption/distribution/metabolism) and measurement accuracy. Here it is clear that the performance of HTPBTK with parameters for a random chemical while worse, is not a marked departure from the performance HTPBTK with the correct parameters or a QSPR model. In second panel we note that at early time points all methods are more accurate than for all time points. The early absorption and distribution phases are dominated by the accuracy of volume of distribution predictions which largely depending on physico-chemical properties (which have not been randomized) and weakly on fup. . We note that the same absorption rate is used for all QSPR models (and indeed for all chemicals) based on the mean absorption observed in Wambaugh et al. [26]. In the third panel of Figure 5 we see that all predictors perform worse in the elimination phase, which is driven by the estimated Clint. For the late time points the specific values of the HTTK in vitro parameters (measured or predicted) have greater influence on the accuracy of the predictions – y-randomization performs notably worse.

Early time points are dominated by the ability to correctly predict peak plasma concentration (Cmax). In Figure 6 we examine each methods accuracy in predicting Cmax as determined from the CvT data. error is always greatest at the low end of the Cmax range

In Figure 7 we examine the ability to predict time-integrated area under the plasma concentration time course (Area Under the Curve or AUC). Again, the empirical fits give a clear best-case scenario, but here the y-randomization more clearly gives a worst-case scenario. The other various methods are all roughly comparable, although again ADMET outperforms even the in vitro measured values.

In Figure 8 we plot each chemical-specific RMSLE by prediction method with each method as a row and each chemical as a column. These data are also provided in Supplemental Table 7. Chemicals and methods have been clustered based upon Euclidean distance. We see that the largest RMSLE is for the Tamoxifen measured data, potentially influencing the performance of in vitro measured data relative to the predictions of the QSPRs.

## Level 3 Analysis

We then proceed on to the third level of evaluation, in which we use the QSPR predictions to predict toxicokinetic summary parameters – volume of distribution (Vd), half-life for elimination from the body (thalf), and whole-body clearance (Cltot) – and compare the predictions to the values estimated from the empirical fits to the CvT data. The values predicted for each method are provided in Supplemental Table 8.

In Figure 9 we examine predicted vs. observed thalf. As summarized in X none of the models are very successful – the highest coefficient of variation is 0.15 for IFS-QSAR, while QSARINS-Chem and HTTK with ADMET both had an R2 of 0.11. The RMSLE for all models, including HTTK with y-randomized data, was just above 1 (a factor of 10x).

The models were distinctly better than the y-randomization for predicting Vd. As shown in Figure 10 and summarized in X , the HTTK algorithm for predicting Vd [34] when used with y-randomized data had no skill. The models performed similarly to the measured *in vitro* data when used with the Vd algorithm – R2 ranged from 0.10 to 0.16 with the measured data being the worst. For all models the RMSLE again indicated a factor of 10x.

In Figure 11 we examine predictions for CLtot, which depends on both elimination rate (inverse of thalf) and Vd. As summarized in X the y-randomized predictions reassuringly have no skill at predicting *in vivo* clearance, while the combination of HTTK and ADMET predicted values had the most, with a R2 of 0.32 and a RMSLE indicating a factor of 17x. Both IFS-QSAR and QSARINS-Chem had comparable R2 of 0.25 and 0.2 (respectively) and an RMSLE indicating a factor of ~20x. The other QSPRs performed about as well as using the *in vitro* measured data.

# Discussion

TK information, including half-life (thalf) and toxicological dose metrics like peak concentration (Cmax) and time-integrated plasma concentration (AUC) are critical for understanding chemical risk. KUnfortunately, TK is chemical-specific, requiring some sort of method tailed or a per chemical basis. L exist . Multiple governments have recognized [42-45] that high throughput (chemical-agnostic) TK models parameterized with chemical-specific *in vitro* data are a powerful tool for facilitating next generation risk assessment based on in vitro screening for chemical toxicity. Governments and industry are continuing to accumulate chemical-specific TK data including both *in vivo* concentration vs. time data in key tissues [27,26] and *in vitro* HTTK data [8,9,11,12,10,19]. However, several thousand chemicals remain in need of TK info; the QSPRs evaluated here provide options to fill this gap.

Since the *in vitro* data for HTTK are limited in number, here we have conducted a collaborative trial of four QSPRs for *in vitro* TK parameters and two additional predictors of *in vivo* TK half-life. The focus of the analysis was using *in vivo* data from rat or human TK studies collected by CvTdb [27]. We aimed to characterize the accuracy of HTTK approaches for new chemicals based on structure-based *in silico* predictions. For novel compounds a consensus quantitative structure-property-relationship (QSPR) approach may yield a reasonable result. This collaborative trial used a database of *in vivo* measured toxicokinetic data to evaluate *in silico* approaches. In vivo data had to be carefully reviewed to include only those data that could be well-described by empirical TK models. Six different sets of QSPR tools for predicting TK were evaluated. Four of the QSPR models made predictions for chemical-specific *in vitro* measurements (HTTK data). An additional consensus prediction was constructed from the various QSPRs. By comparing predictions with observations, the root mean squared log10 error (RMSLE) and other key statistics could be calculated. The RMSLE characterizes the expected accuracy for new predictions. Given sufficient observations for evaluation, the RMSLE can be interpreted as a coefficient of variation for normally distributed errors about the prediction. For normally distributed errors one has 95% confidence that the actual value will occur within +- 2 RMSLE of the prediction.

Throughout this effort we have reported statistics on different subsets of chemicals. First, HTTK in vitro data were not available for all the chemicals with CVT data. For the context of decision makers relying either on in vitro HTTK or QSPR-based HTTK we report chiefly on the statistics associated with either the in vitro only chemicals or the chemicals with no existing in vitro data.

QSPRs were first evaluated for their ability to predict the parameters fup and Clint expected by high throughput PBTK models. These in vitro parameters, of course, are not true TK. Rather they represent rapid method rapid methods for partially characterizing TK. [18,15]. However, there is an ongoing proliferation of high throughput PBTK models developed to make use of these in vitro parameters to allow HT-PBTK models to make chemical-specific predictions [46,30,15,47,48]. Most QSPRs predict plasma binding well (RMSLE 0.03 – 0.07) but have difficulty predicting metabolic clearance (RMSLE 0.37 – 1.28). Even if qsar perfectly reproduces in vitro, only as good as in vitro

Predictions for the full concentration time-course were also evaluated. Our “level 2” evaluation used the QSPR predicted parameter values within a physiologically based TK (PBTK) model to predict the full TK chemical concentration time course. To frame the evaluation, statistics were also calculated for empirical TK model fits to the in vivo data and HTTK based on randomly selected (incorrect) chemical parameters. The in vivo fits were intended to approximate a best case given that the data themselves are noisy. The y-randomization approximates a worst case, taking into account potential correlations within the chemical data. PBTK models parameterized with QSPRS perform closer to y-randomized predictions than to empirical fits to the data. First, although nearly 100 chemicals is substantial, the existing annotated TK data do not constitute machine learning “Big Data” which might rely on thousands if not millions of observations [49]. Although more than one thousand chemicals have available in vitro HTTK measurements, it is still a constrained random set of chemicals reflecting the correlation and the property distribution of the set. For example, to be suitable for *in vitro* measurement the volatility and solubility of the chemicals must be somewhat constrained. When chemical space is too narrow the overall statistics are limited; with the CvTdb, like any data set, one can only evaluate and model things that vary across our dataset. Among the evaluation chemicals only two had metabolic clearance (Clint above 103 µL/min/106 hepatocytes and 93% are within two-fold of the median. The two parameters fup and Clint further interact in how they influence TK; if a chemical has low metabolic clearance, it may accumulate regardless of how highly the chemical binds; conversely if a chemical is highly bound it may not matter how fast the free chemical clears.

Consensus is not so much the particular models as multiple ways of predicting metabolism – if any predict clearance chemical may be moe likely to be metabolized

5 out of 67 chemicals with CvT data have no measured in vitro clearance compared with 254 out of 1023 measured Clint values ( 25 percent).

Here we have found the HTTK PBTK model performed similarly when using TK QSPRs for Clint and fup as when the actual in vitro measured data were used. Fits to the in vivo data had RMSLE of 0.62, while predictions based on *in vitro* HTTK had RMSLE of 1.18 without using any *in vivo* data. RMSLE for individual QSPRs ranged from 1.19 to 1.3. Any one QSPR performed slightly better than using *in vitro* values for random chemicals (RMSLE 1.35). A consensus prediction using the maximum clearance predicted across all QSPRs had RMSLE 1.09 for the full time course – this is better than using the *in vitro* measured data for the evaluation chemicals.

IVBP over predicts in vitro clearance, dragging up predictions (correctly as it turns ouyt)

* Looking at Fig 4, the models tend to underpredict blood levels compared to the in vivo data. For IVIVE, we are trying to predict the oral dose that results in an equivalent blood level to the in vitro. If the HTTK models tend to underpredict blood levels compared to in vivo, then their oral equivalent dose would be higher than that obtained by in vivo studies.
  + Possible reasons: either clearing the chemical too quickly, or our oral absorption is not correct (too slow?)
* discussion on why they are all performing similarly in late time points vs differences when you consider overall trends and early trends might be useful.

Among the summary TK statistics (level 3) evaluated, the related quantities AUC, CLtot, and half-life are more challenging to predict. Vd only depends on partitioning (partially characterized by fup) and Cmax only depends on Vd; neither of these two quantities depend upon Clint. For AUC, optimized fits to the *in vivo* data indicated a RMSLE of 0.5 while using *in vitro* values for random (incorrect) chemicals had an AUC RMSLE of 0.89. Using chemical-specific *in vitro* data predicted AUC with RMSLE 0.74, while QSPRs ranged in accuracy from 0.74 to 0.84. The consensus prediction using the maximum clearance predicted across all QSPRs predicted AUC with RMSLE 1.03 – this is better than using the *in vitro* measured data for the evaluation chemicals. For total clearance CLtot the RMSLE 1.57 and R2 0.03.

For other TK summary statistics (level 3) we observed Cmax *in vivo* fits gave RMSLE 0.5 and random chemicals gave 1.7. Chemical-specific *in vitro* data had RMSLE 1.3 and QSPRs ranged from 1.3-1.4.

Vd had RMSLE 0.89 with R2 0.04.

Comparison with previous evaluation:

We report the in vitro-based HTTK RMSLE for the full concentration time course (Cvt) data (“level 2” analysis) for 83 chemicals as X.

Wambaugh et al. [26] found evaluated HTTK-based IVIVE for summary TK endpoints (what we call “level 3” here) for TK using just over forty chemicals. The mean squared error (MSE) observed for Vd was 4.4. For CLtot the MSE was 2.4 for pharma, and 2.93 for non-pharma. However, the fraction of variance explained (R2) was 0.19 for pharma, 0.5 for non-pharma. For Cmax Wambaugh et al. [26] found MSE 5, R2 0.48. Finally, for AUC the MSE was 3.8, with R2 0.62. Why the drop-off?

note that CLtot and AUCinfinity are related

Limitations of model for in vitro intrinsic clearance

* admet is five cyps -- not an apples to apples comparison (figure 2)
* what are impacts of 3d effects in qsar? is that something we're missing?
* all the models are 2d right now because we don't have data to train qsar models to predict 3d differences (for example chiral pairs)
* - everything is trained on human
* What is breakdown rat vs. human in evaluation set?
* We do not expect Fabs to correlate between rat and human – good reason not to worry about Caco2 here
* - only vary in HTPBTK here by physiology – using human clint and fup

Discuss chemicals with worst RSMLEs

* Eventually ?

These QSPRs will enable public health risk-based prioritization of many more chemicals in commerce and the environment than in vivo and in vitro testing alone.

# Acknowledgements

The authors thank Drs. X and Y for their helpful U.S. EPA internal reviews of the manuscript.

# Conflict of Interest

Please declare any COI here

# Funding

Employees of the United States Environmental Protection Agency (EPA) were funded through EPA’s Office of Research and Development (ORD). The Oak Ridge Institute for Science

and Education provided funding for C. Cook. The National Institute of Environmental Health Sciences provided funding for S. Ferguson and K. Mansouri.

Please add your funding here

The views expressed in this publication are those of the authors and do not necessarily represent the views or policies of the U.S. EPA. Reference to commercial products or services does not constituteendorsement.

# References

1. O'Flaherty EJ (1981) Toxicants and drugs: kinetics and dynamics. John Wiley & Sons,

2. Coecke S, Pelkonen O, Leite SB, Bernauer U, Bessems JG, Bois FY, Gundert-Remy U, Loizou G, Testai E, Zaldívar J-M (2013) Toxicokinetics as a key to the integrated toxicity risk assessment based primarily on non-animal approaches. Toxicology in Vitro 27 (5):1570-1577

3. Sobus JR, Tan Y-M, Pleil JD, Sheldon LS (2011) A biomonitoring framework to support exposure and risk assessments. Science of the Total Environment 409 (22):4875-4884

4. National Research Council (1983) Risk Assessment in the Federal Government: Managing the Process. In. National Academies Press, Washington (DC). doi:10.17226/317

5. Wetmore BA (2015) Quantitative in vitro-to-in vivo extrapolation in a high-throughput environment. Toxicology 332:94-101

6. Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, Knudsen T, Martin M, Padilla S, Reif D, Richard AM, Rotroff D, Sipes NS, Dix D (2012) Update on EPA’s ToxCast program: providing high throughput decision support tools for chemical risk management. Chemical Research in Toxicology 25 (7):1287-1302

7. Tice RR, Austin CP, Kavlock RJ, Bucher JR (2013) Improving the human hazard characterization of chemicals: a Tox21 update. Environmental Health Perspectives 121 (7):756

8. Rotroff DM, Wetmore BA, Dix DJ, Ferguson SS, Clewell HJ, Houck KA, LeCluyse EL, Andersen ME, Judson RS, Smith CM, Sochaski MA, Kavlock RJ, Boellmann F, Martin MT, Reif DM, Wambaugh JF, Thomas RS (2010) Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. Toxicological Sciences 117 (2):348-358

9. Wetmore BA, Wambaugh JF, Ferguson SS, Sochaski MA, Rotroff DM, Freeman K, Clewell HJ, 3rd, Dix DJ, Andersen ME, Houck KA, Allen B, Judson RS, Singh R, Kavlock RJ, Richard AM, Thomas RS (2012) Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. Toxicological Sciences 125 (1):157-174. doi:10.1093/toxsci/kfr254

10. Tonnelier A, Coecke S, Zaldívar J-M (2012) Screening of chemicals for human bioaccumulative potential with a physiologically based toxicokinetic model. Archives of Toxicology 86 (3):393-403

11. Wetmore BA, Wambaugh JF, Allen B, Ferguson SS, Sochaski MA, Setzer RW, Houck KA, Strope CL, Cantwell K, Judson RS, LeCluyse E, Clewell HJ, Thomas RS, Andersen ME (2015) Incorporating High-Throughput Exposure Predictions With Dosimetry-Adjusted In Vitro Bioactivity to Inform Chemical Toxicity Testing. Toxicological Sciences 148 (1):121-136. doi:10.1093/toxsci/kfv171

12. Wambaugh JF, Wetmore BA, Ring CL, Nicolas CI, Pearce R, Honda G, Dinallo R, Angus D, Gilbert J, Sierra T, Badrinarayanan A, Snodgrass B, Brockman A, Strock C, Setzer W, Thomas RS (2019) Assessing Toxicokinetic Uncertainty and Variability in Risk Prioritization. Toxicological Sciences 172:235-251

13. Bell SM, Chang X, Wambaugh JF, Allen DG, Bartels M, Brouwer KLR, Casey WM, Choksi N, Ferguson SS, Fraczkiewicz G, Jarabek AM, Ke A, Lumen A, Lynn SG, Paini A, Price PS, Ring C, Simon TW, Sipes NS, Sprankle CS, Strickland J, Troutman J, Wetmore BA, Kleinstreuer NC (2018) In vitro to in vivo extrapolation for high throughput prioritization and decision making. Toxicology in Vitro 47:213-227. doi:<https://doi.org/10.1016/j.tiv.2017.11.016>

14. Kavlock RJ, Bahadori T, Barton-Maclaren TS, Gwinn MR, Rasenberg M, Thomas RS (2018) Accelerating the Pace of Chemical Risk Assessment. Chemical Research in Toxicology 31 (5):287-290. doi:10.1021/acs.chemrestox.7b00339

15. Breen M, Ring CL, Kreutz A, Goldsmith M-R, Wambaugh JF (2021) High-throughput PBTK models for in vitro to in vivo extrapolation. Expert Opinion on Drug Metabolism & Toxicology 17 (8):903-921

16. Armitage JM, Hughes L, Sangion A, Arnot JA (2021) Development and intercomparison of single and multicompartment physiologically-based toxicokinetic models: Implications for model selection and tiered modeling frameworks. Environment International 154:106557

17. US EPA (2019) Directive to Prioritize Efforts to Reduce Animal Testing. Washington D.C.

18. Wang Y-H (2010) Confidence Assessment of the Simcyp Time-Based Approach and a Static Mathematical Model in Predicting Clinical Drug-Drug Interactions for Mechanism-Based CYP3A Inhibitors. Drug Metabolism and Disposition 38 (7):1094-1104. doi:10.1124/dmd.110.032177

19. Paini A, Cole T, Meinero M, Carpi D, Deceuninck P, Macko P, Palosaari T, Sund J, Worth A, Whelan M (2020) EURL ECVAM in vitro hepatocyte clearance and blood plasma protein binding dataset for 77 chemicals.

20. National Academies of Sciences E, Medicine (2017) Using 21st century science to improve risk-related evaluations. National Academies Press,

21. Arnot JA, Brown TN, Wania F (2014) Estimating screening-level organic chemical half-lives in humans. Environmental science & technology 48 (1):723-730

22. Papa E, Sangion A, Arnot JA, Gramatica P (2018) Development of human biotransformation QSARs and application for PBT assessment refinement. Food and Chemical Toxicology 112:535-543

23. Sipes NS, Wambaugh JF, Pearce R, Auerbach SS, Wetmore BA, Hsieh J-H, Shapiro AJ, Svoboda D, DeVito MJ, Ferguson SS (2017) An Intuitive Approach for Predicting Potential Human Health Risk with the Tox21 10k Library. Environmental Science & Technology 51 (18):10786-10796. doi:10.1021/acs.est.7b00650

24. Pradeep P, Patlewicz G, Pearce R, Wambaugh J, Wetmore B, Judson R (2020) Using chemical structure information to develop predictive models for in vitro toxicokinetic parameters to inform high-throughput risk-assessment. Computational Toxicology 16:100136

25. Dawson DE, Ingle BL, Phillips KA, Nichols JW, Wambaugh JF, Tornero-Velez R (2021) Designing QSARs for Parameters of High-Throughput Toxicokinetic Models Using Open-Source Descriptors. Environmental science & technology 55 (9):6505-6517

26. 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 (2018) Evaluating in vitro-in vivo extrapolation of toxicokinetics. Toxicological Sciences 163 (1):152-169

27. Sayre RR, Wambaugh JF, Grulke CM (2020) Database of pharmacokinetic time-series data and parameters for 144 environmental chemicals. Scientific Data

28. Shibata Y, Takahashi H, Chiba M, Ishii Y (2002) Prediction of hepatic clearance and availability by cryopreserved human hepatocytes: an application of serum incubation method. Drug Metabolism and Disposition 30 (8):892-896

29. Waters NJ, Jones R, Williams G, Sohal B (2008) Validation of a rapid equilibrium dialysis approach for the measurement of plasma protein binding. Journal of Pharmaceutical Sciences 97 (10):4586-4595

30. Pearce RG, Setzer RW, Strope CL, Sipes NS, Wambaugh JF (2017) httk: R package for high-throughput toxicokinetics. Journal of Statistical Software 79 (1):1-26

31. R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

32. Baumer B, Udwin D (2015) R markdown. Wiley Interdisciplinary Reviews: Computational Statistics 7 (3):167-177

33. Schmitt W (2008) General approach for the calculation of tissue to plasma partition coefficients. Toxicology in Vitro 22 (2):457-467

34. Pearce RG, Setzer RW, Davis JL, Wambaugh JF (2017) Evaluation and calibration of high-throughput predictions of chemical distribution to tissues. Journal of Pharmacokinetics and Pharmacodynamics 44 (6):549-565. doi:10.1007/s10928-017-9548-7

35. Davies B, Morris T (1993) Physiological parameters in laboratory animals and humans. Pharmaceutical Research 10 (7):1093-1095

36. Ruark CD, Hack CE, Robinson PJ, Mahle DA, Gearhart JM (2014) Predicting passive and active tissue: plasma partition coefficients: interindividual and interspecies variability. Journal of pharmaceutical sciences 103 (7):2189-2198

37. Akaike H (1998) Information theory and an extension of the maximum likelihood principle. In: Selected papers of hirotugu akaike. Springer, pp 199-213

38. Schmidt CW (2016) TSCA 2.0: A new era in chemical risk management. National Institute of Environmental Health Sciences,

39. Patlewicz G, Richard AM, Williams AJ, Grulke CM, Sams R, Lambert J, Noyes PD, DeVito MJ, Hines RN, Strynar M (2019) A chemical category-based prioritization approach for selecting 75 per-and polyfluoroalkyl substances (PFAS) for tiered toxicity and toxicokinetic testing. Environmental health perspectives 127 (01):014501

40. U.S. Centers for Disease Control and Prevention (2012) National Health and Nutrition Examination Survey. <http://www.cdc.gov/nchs/nhanes.htm>. Accessed 07/15/2013

41. Wang A, Gerona RR, Schwartz JM, Lin T, Sirota M, Morello-Frosch R, Woodruff TJ (2018) A suspect screening method for characterizing multiple chemical exposures among a demographically diverse population of pregnant women in San Francisco. Environmental Health Perspectives

126 (7):077009

42. National Academies of Sciences E, and Medicine, (2017) Using 21st Century Science to Improve Risk-Related Evaluations. The National Academies Press, Washington, DC. doi:doi:10.17226/24635

43. U.S. Environmental Protection Agency (2021) U.S. EPA. A Proof-of-Concept Case Study Integrating Publicly Available Information to Screen Candidates for Chemical Prioritization under TSCA. Washington, DC. doi:<https://doi.org/10.23645/epacomptox.14878125>

44. Health Canada (2021) Bioactivity Exposure Ratio: Application in Priority Setting and Risk Assessment. Ottawa, ON, Canada

45. Paini A, Cole T, Meinero M, Carpi D, Deceuninck P, Macko P, Palosaari T, Sund J, Worth A, Whelan M (2020) EURL ECVAM in vitro hepatocyte clearance and blood plasma protein binding dataset for 77 chemicals.

46. Jamei M, Marciniak S, Feng K, Barnett A, Tucker G, Rostami-Hodjegan A (2009) The Simcyp® population-based ADME simulator. Expert Opinion on Drug Metabolism & Toxicology 5 (2):211-223

47. Linakis MW, Sayre RR, Pearce RG, Sfeir MA, Sipes NS, Pangburn HA, Gearhart JM, Wambaugh JF (2020) Development and Evaluation of a High Throughput Inhalation Model for Organic Chemicals Journal of Exposure Science & Environmental Epidemiology 30:866-877

48. Bernstein AS, Kapraun DF, Schlosser PM (2021) A model template approach for rapid evaluation and application of physiologically based pharmacokinetic models for use in human health risk assessments: a case study on per-and polyfluoroalkyl substances. Toxicological Sciences

49. Kitchin R, McArdle G (2016) What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society 3 (1):2053951716631130

50. Mansouri K, Grulke CM, Judson RS, Williams AJ (2018) OPERA models for predicting physicochemical properties and environmental fate endpoints. Journal of Cheminformatics 10 (1):10

51. Mansouri K, Chang X, Allen D, Judson R, Williams A, Kleinstreuer N (2021) OPERA models for ADME properties and toxicity endpoint. Paper presented at the Society of Toxicology Annual Meeting,

# Figures

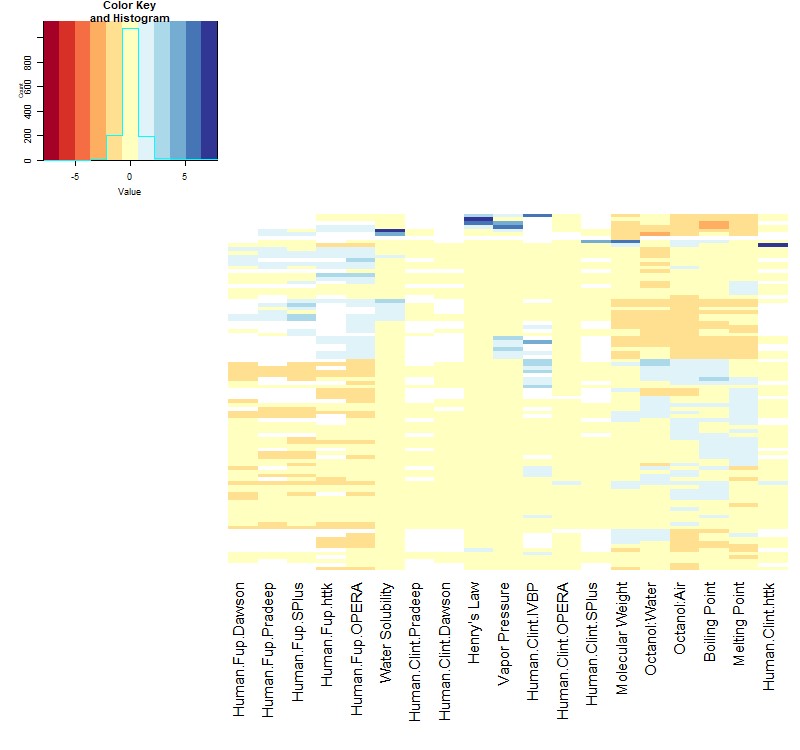


Figure :Columns in this heatmap indicate the physico chemical properties and measured/predicted values for in vitro TK (intrinsic hepatic clearance, Clint, and fraction unbound in plasma, fup). Each row corresponds to one of the 83 chemicals. The in vitro TK measurements (“Human.Clint.InVitro” and “Human.Fup.InvItro”) and predictions for these values from the various QSPRs (Table 2) are indicated by name. Data are normalized on a per column basis by centering (subtracting the mean) and scaling (by standard deviation). Thus, the "Value" of each entry in heatmap indicates the number of standard deviations from the mean. Blank values indicate no prediction.

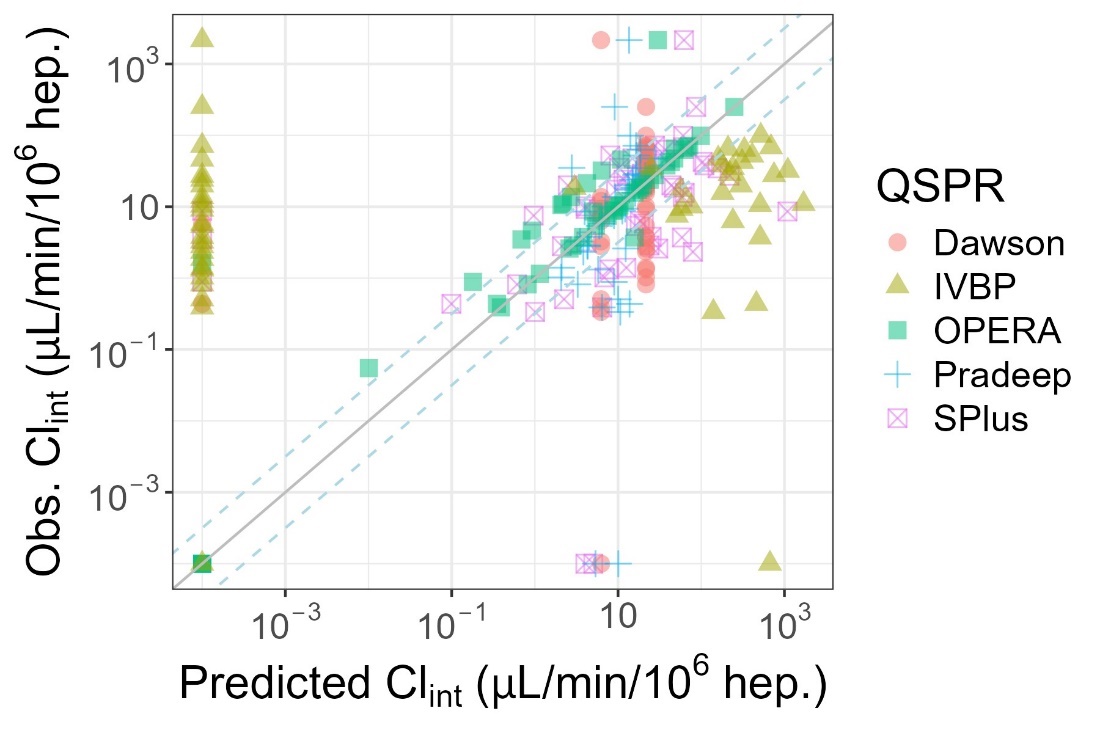
\ 

Figure : Evaluation of Predictions for Intrinsic Hepatic Clearance (Clint). Zero values were plotted at 10-1, the solid line indicates identity (1:1) while the dashed lines indicate 3.2-fold difference.



Figure : Evaluation of Predictions for Fraction Unbound in Plasma (fup). Zero values were plotted at 10-4, the solid line indicates identity (1:1) while the dashed lines indicate 3.2-fold difference

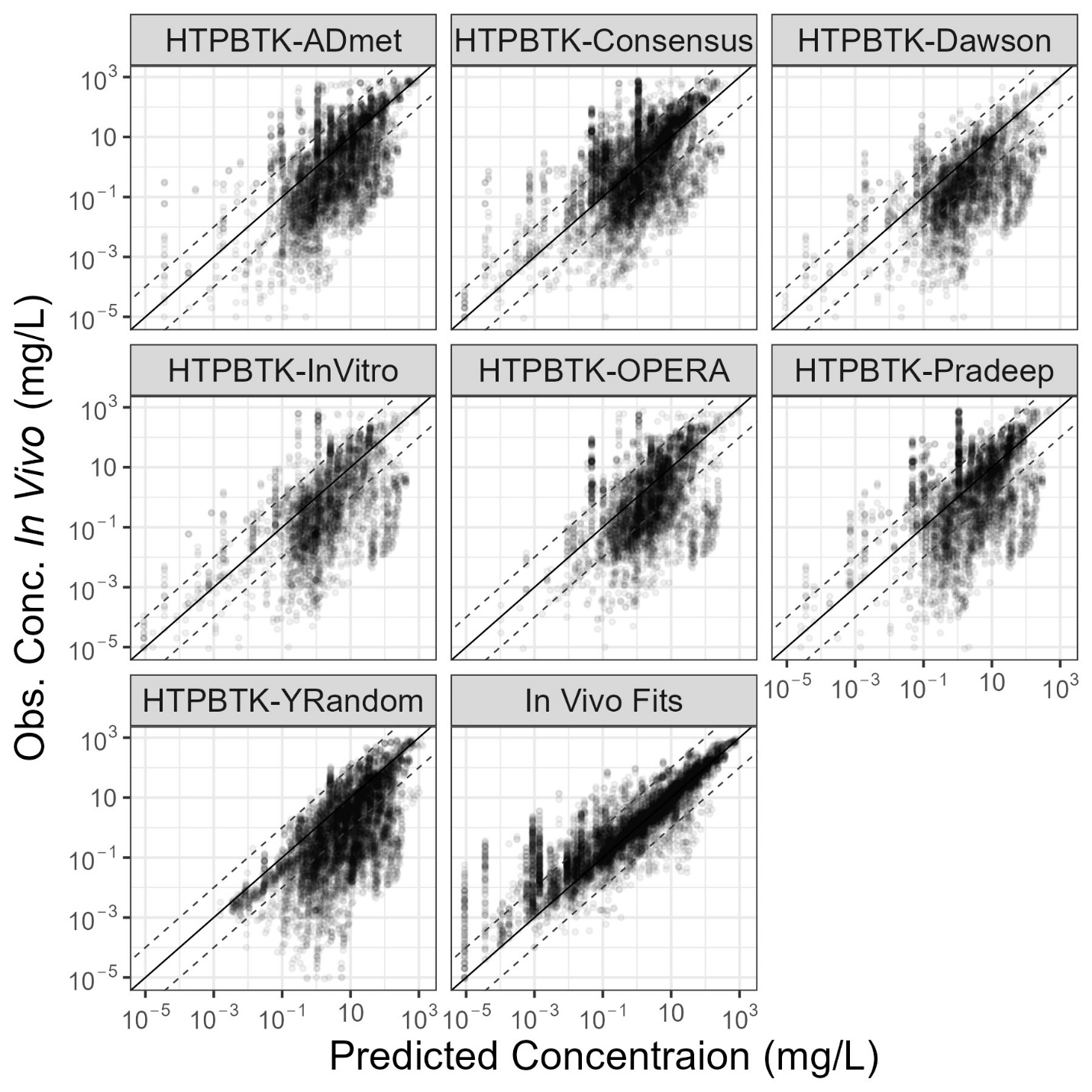


Figure : Comparison of in vivo measured chemical concentration vs. time (CvT) data [27] vs. predictions for empirical models fit to the data (“In Vivo Fits”), and predictions for a PBTK model (“HTTK”) parameterized with chemical specific values either measured in vitro (“HTTK-InVitro”) or predicted with various QSPRs. In each sub-plot the y-axis shows the measured data while the x-axis shows the predictions made using chemical-specific parameters from the various sources. The solid line indicates identity (1:1) while the dashed lines indicate ten-fold difference.

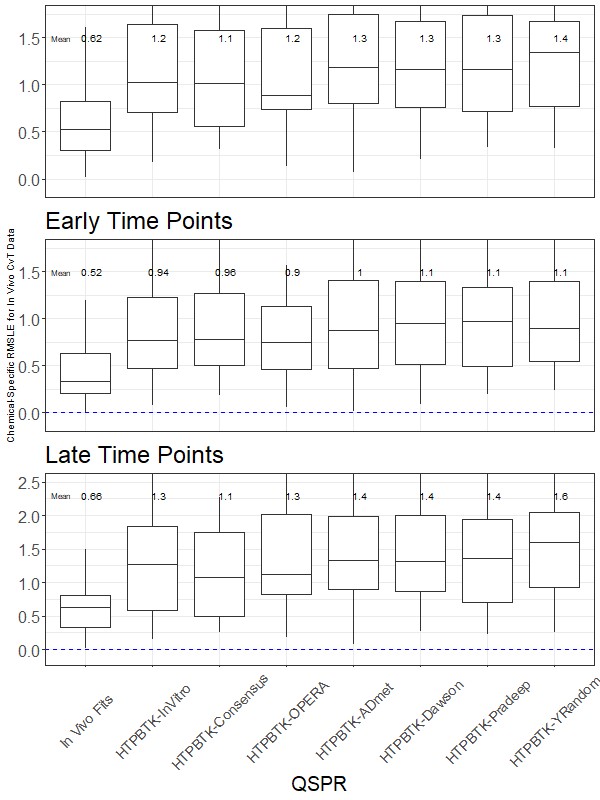


Figure : Chemical-Specific RMSLE for In vivo CvT Data. The upper and lower extent of the box for each model indicates the 25th to 75th quantiles, the mid-line indicates the median (50th quantile) and vertical line indicates 1.5x the range of the box.

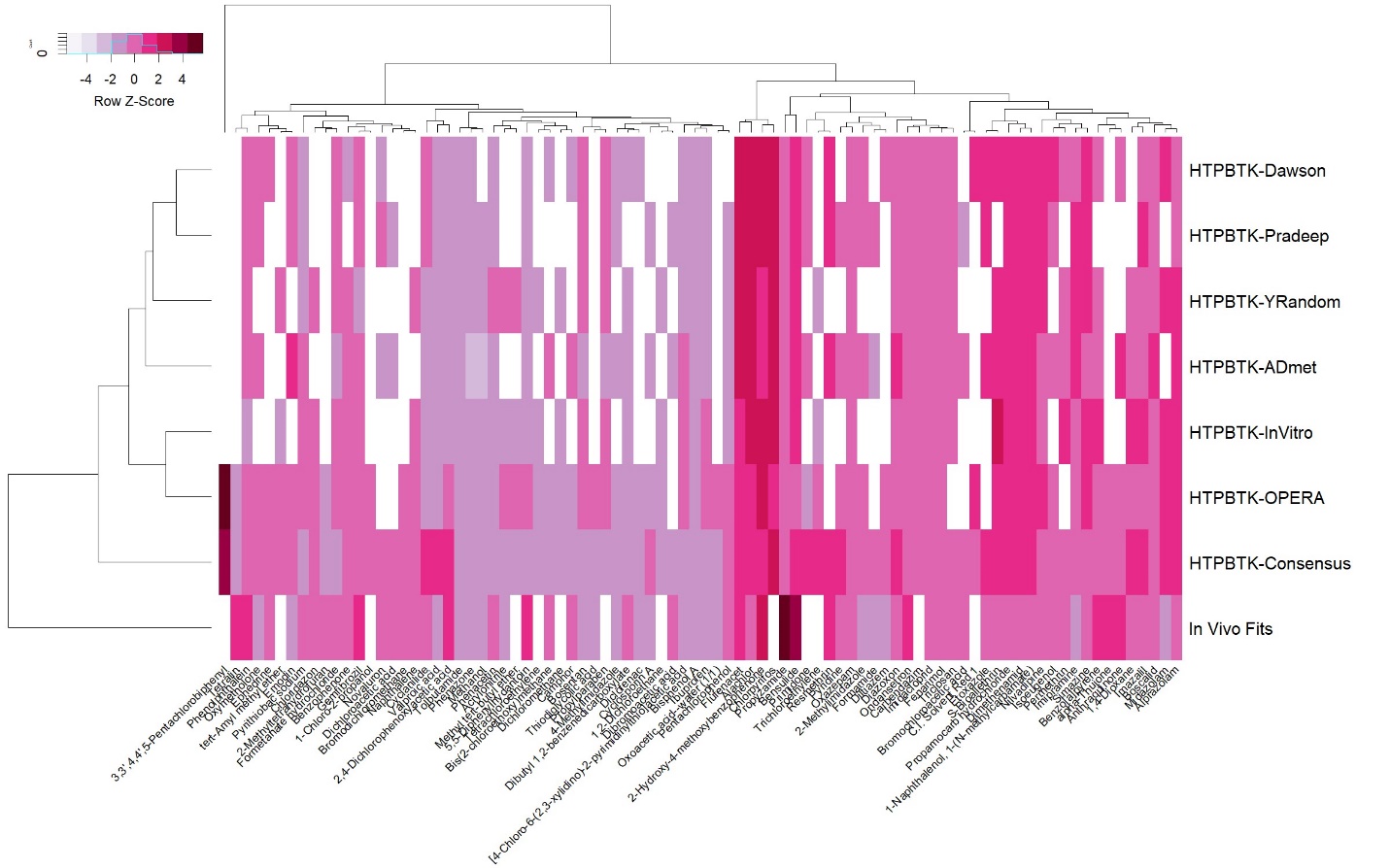
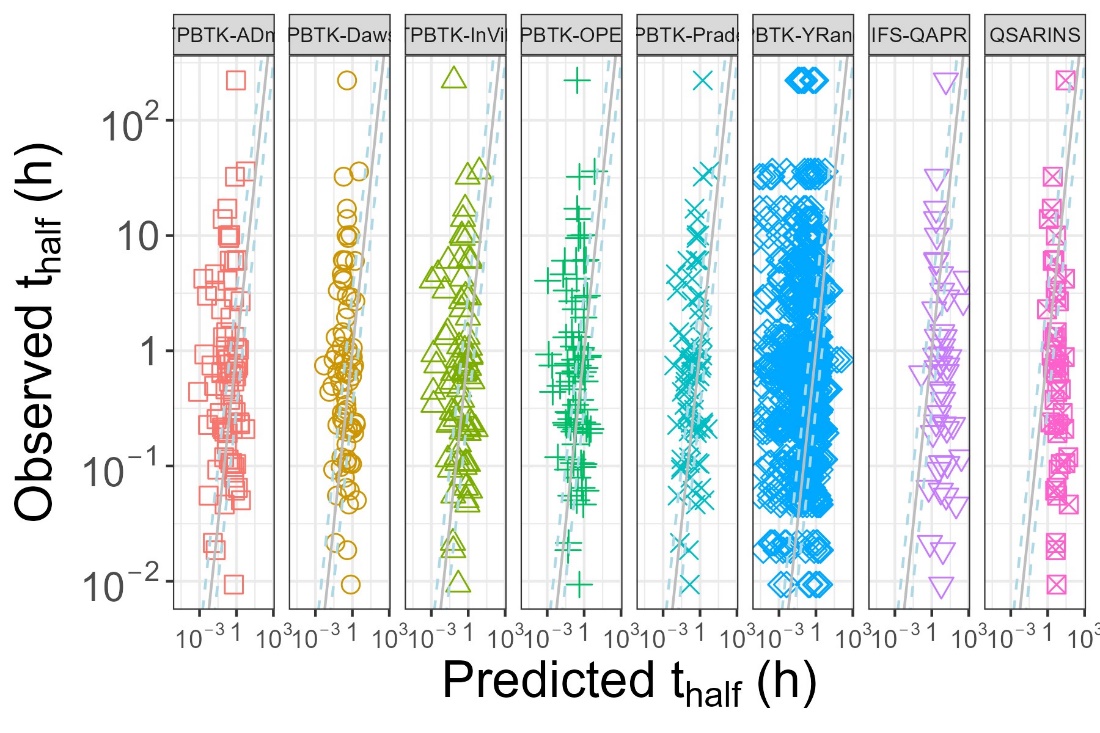


Figure : Values in this heatmap are the root mean squared log10 error (RMSLE) by chemical and predictor (that is, lighter indicates better predictive ability). The columns indicate different evaluation chemicals. The rows give the different prediction methods: The empirical fits to the data are given by “FitsToData”. All other values are calculated using the HTTK PBTK model and either measured values “In vitro”, y-randomized measured values (“Y-Random”) or the various QSPRS.



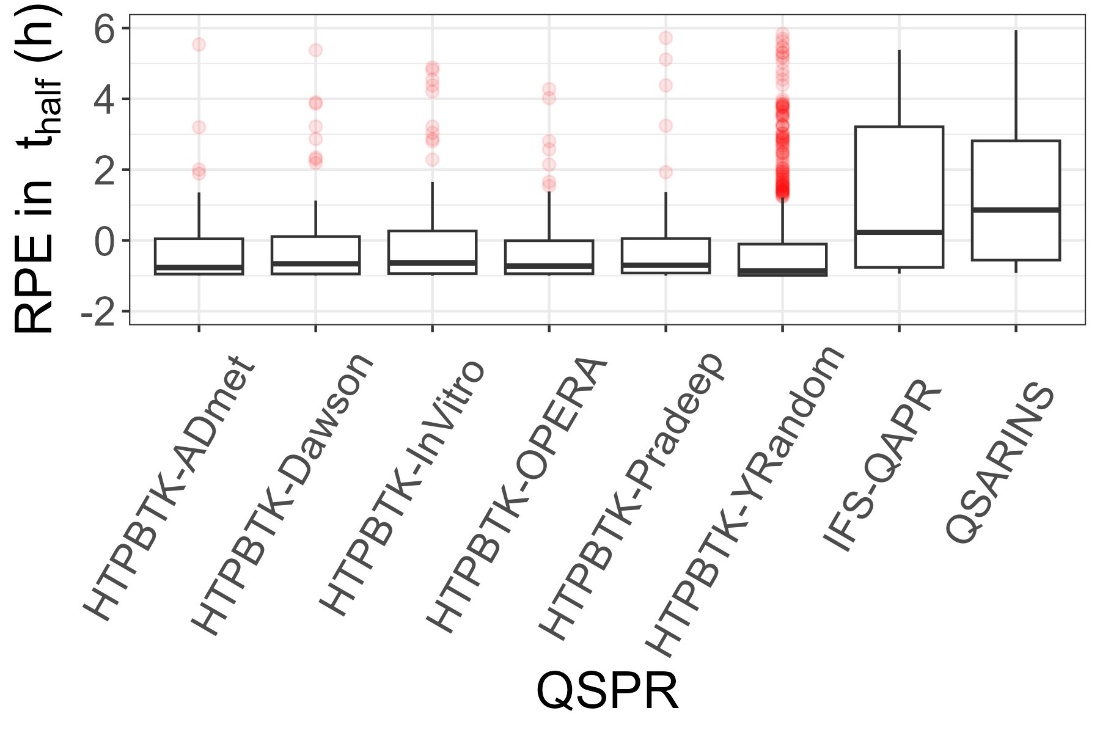
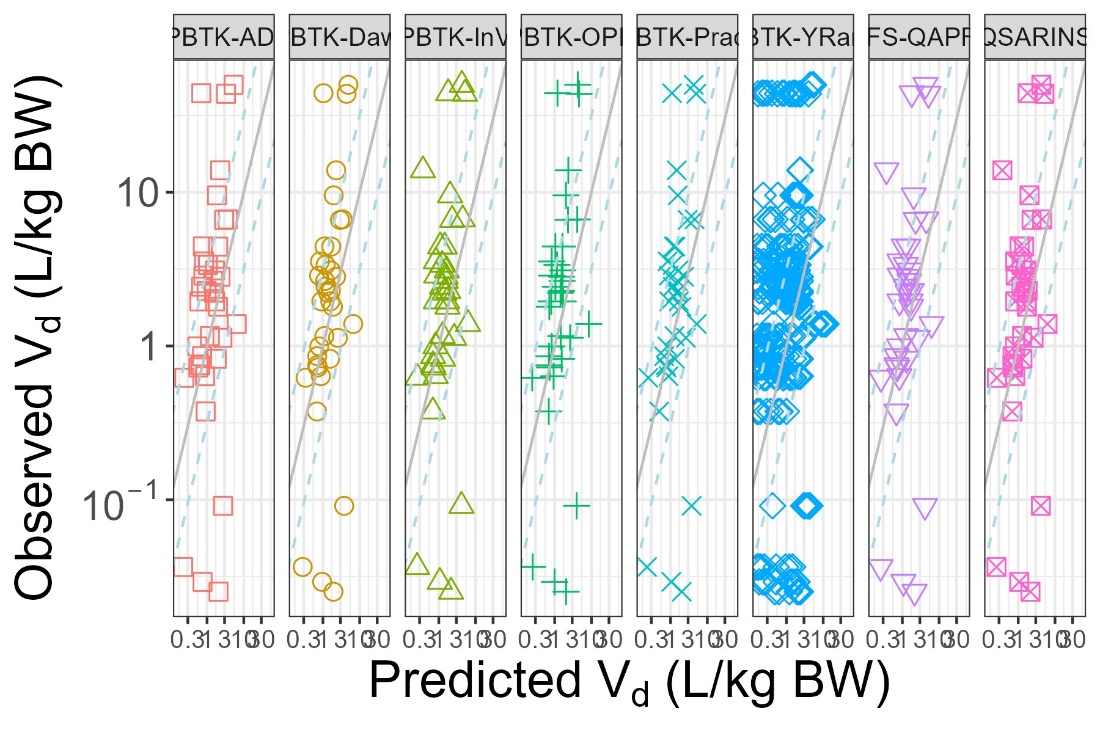


Figure : Comparison of “observed” chemical half-lives based empirical model fits and predictions for chemical half-life based on a PBPK model (“HTTK”) parameterized with chemical specific values either measured *in vitro* (“HTTK-InVitro”) or predicted with various QSPRs. The upper panel shows a scatter plot of predicted vs. observed values, while the lower panel shows the distribution of relative predictive error (RPE). The upper and lower extent of the box for each model indicates the 25th to 75th quantiles, the mid-line indicates the median (50th quantile) and vertical line indicates 1.5x the range of the box.



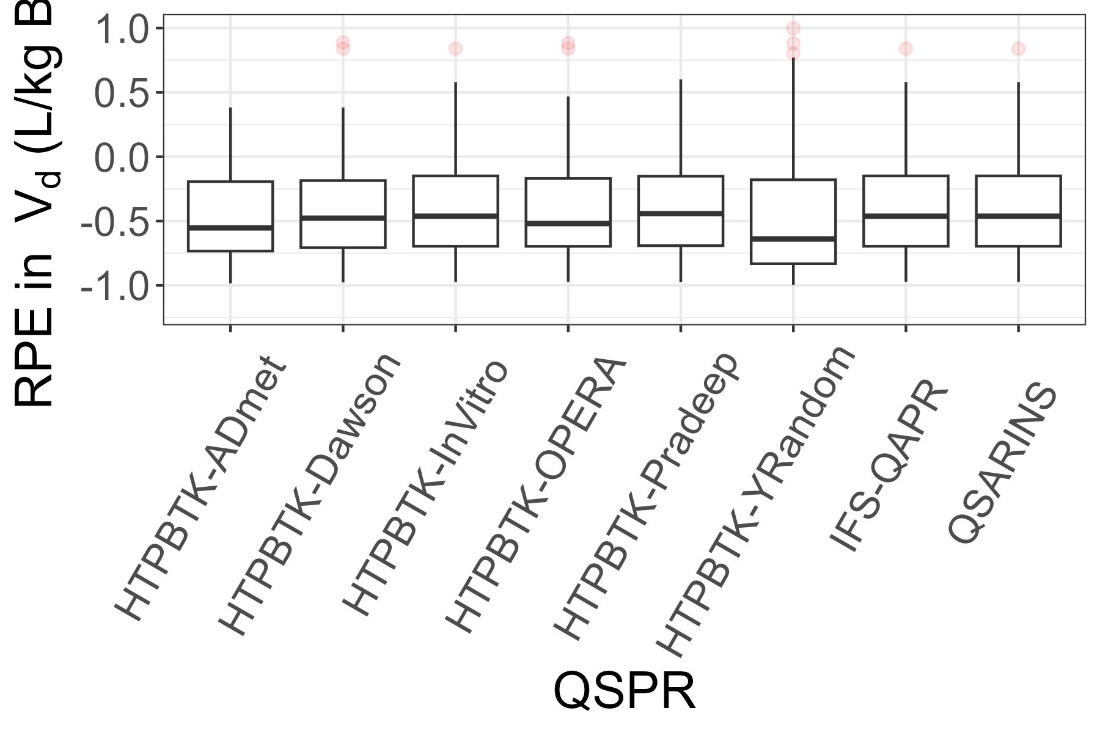
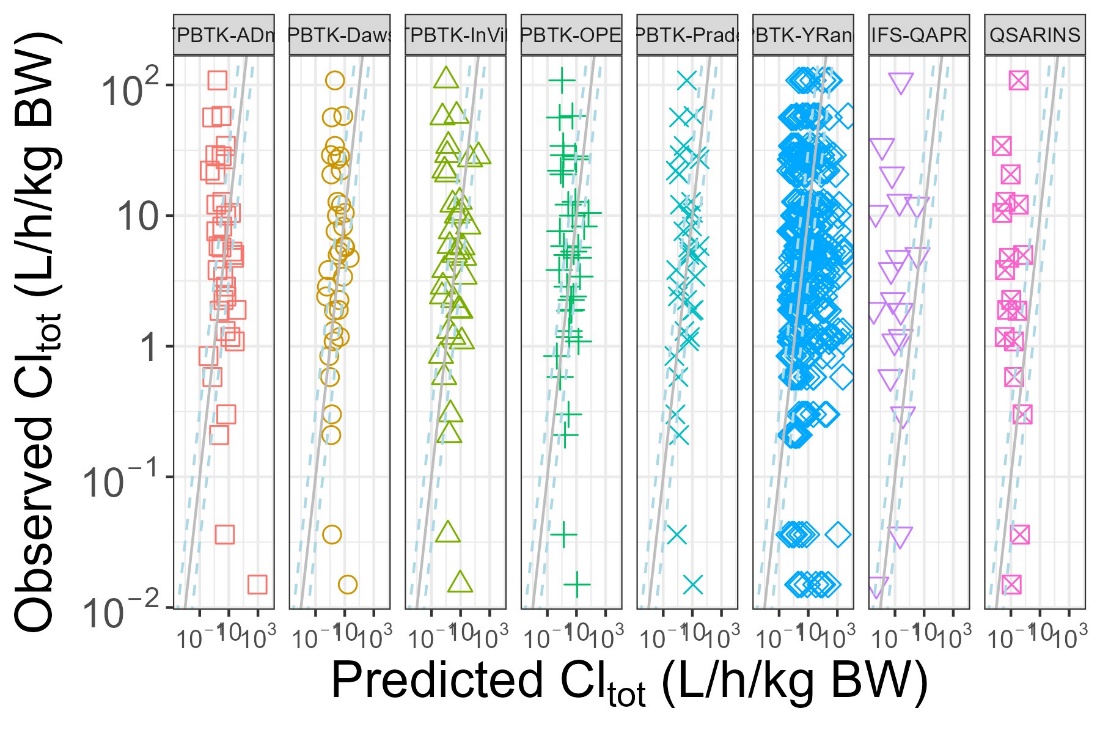


Figure : Comparison of “observed” chemical volumes of distribution based on empirical model fits and predictions for chemical half-life based on a PBPK model (“HTTK”) parameterized with chemical specific values either measured *in vitro* (“HTTK-InVitro”) or predicted with various QSPRs. The upper panel shows a scatter plot of predicted vs. observed values, while the lower panel shows the distribution of relative predictive error (RPE). The upper and lower extent of the box for each model indicates the 25th to 75th quantiles, the mid-line indicates the median (50th quantile) and vertical line indicates 1.5x the range of the box.



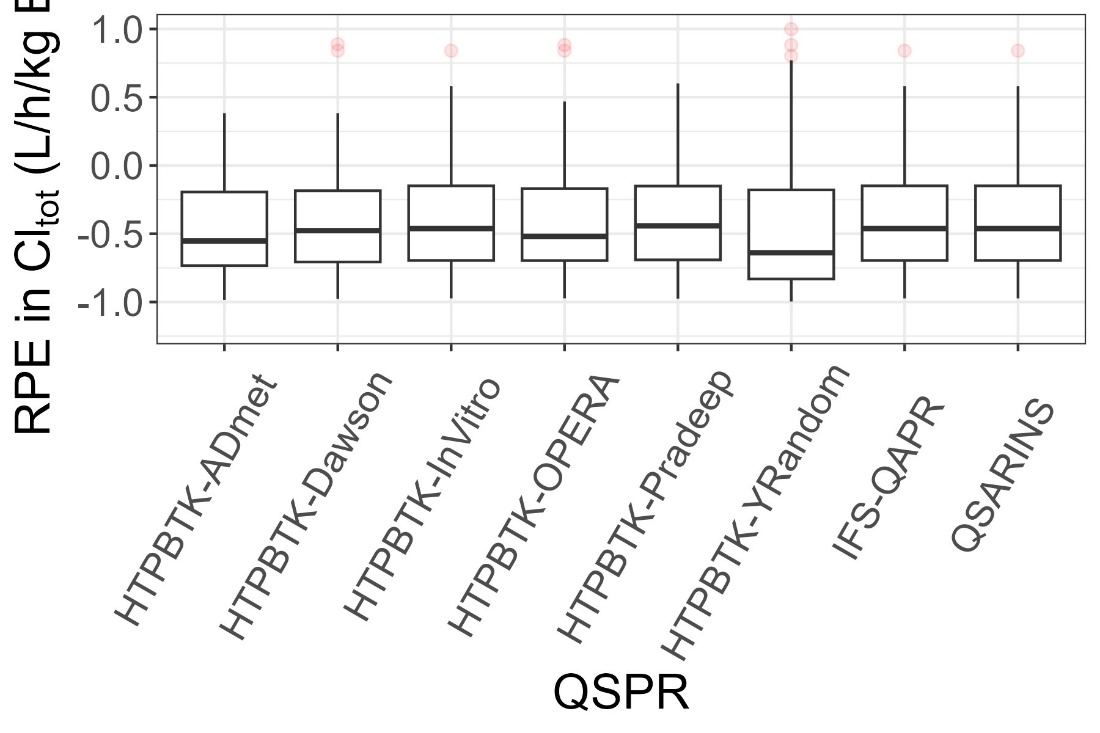


Figure : Comparison of “observed” chemical whole body clearance (Cltot) based on empirical model fits and predictions for chemical half-life based on a PBPK model (“HTTK”) parameterized with chemical specific values either measured *in vitro* (“HTTK-InVitro”) or predicted with various QSPRs. The upper panel shows a scatter plot of predicted vs. observed values, while the lower panel shows the distribution of relative predictive error (RPE). The upper and lower extent of the box for each model indicates the 25th to 75th quantiles, the mid-line indicates the median (50th quantile) and vertical line indicates 1.5x the range of the box.

# Tables

Table Three Levels of Evaluation were performed

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evaluation** | **TK Quantities** | **Chemicals for Evaluation** | **QSPR Models Evaluated** | **Reference** |
| **Level 1** | *In vitro* TK Measurements  (fup, Clint) | 63 with Measured *In vitro* Data | 4 (fup),  5 (Clint) | [30] |
| **Level 2** | TK Concentration vs. Time  (all points, Cmax, time-integral/AUC) | 83 with predictions across multiple QSPRs and empirical model fits | 5 + Consensus | [27] |
| **Level 3** | Summary Statistics  (Vd, thalf, Cltot) | 83 | 6 | [27] |

Table QSPR Models Evaluated

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Predictions** | **Mechanism** | **Reference** |
| Simulations Plus ADMET Predictor® | Level 1  (*in vitro* parameters) | Sum of CYP-specific Artificial Neural Network (ANN) | [23] |
| Pradeep 2020 | Level 1 | Random forest and support vectors method | [24] |
| Dawson 2021 | Level 1 | Random forest, clearance organized by categories | [25] |
| OPERA | Level 1 | Nearest-neighbors | [50,51] |
| IFS-QSAR | Level 3  (Half-lives) | Fragment-based Multiple Linear Regressors (MRL) | [21] |
| QSARINS-Chem | Level 3 | Ordinary Least Squares MLR | [22] |

Table : Biases of the QSPRs for predicting in vitro measured values in terms of log10 fold error (FE) – no bias would be FE = 0



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Num  Clint  Compared | Median  Clint  AbsFE | Median  Clint  FE | Num  Fup  Compared | Median  fup  AbsFE | Median  Fup  FE |
| **SPlus** | 51 | 0.545 | 0.105 | 46 | 0.289 | -0.125 |
| **Dawson** | 49 | 0.376 | -0.00525 | 44 | 0.153 | 0.0264 |
| **Pradeep** | 45 | 0.232 | 0.00985 | 40 | 0.123 | -0.0485 |
| **OPERA** | 46 | 0.00127 | -0.00013 | 41 | 0.03 | 1.27E-05 |
| **IVBP** | 52 | 2.82 | 0.183 | 0 |  |  |

Table 4 Level 2 Analysis of predictions based on empirical model fits and a PBPK model (“HTTK”) parameterized with chemical specific values either measured in vitro (“HTTK-InVitro”) or predicted with various QSPRs. RMSLE was first calculated on a per chemical basis and then averaged across chemicals. The number of chemicals in each calculation is given in parentheses. The four sets of chemicals refer to those with in vitro HTTK parameters (“In Vitro”), those without in vitro HTTK parameters (“No In Vitro”), those chemicals for which all QSPRs could make predictions, and the maximum number of predictions available for each chemical. “HTTK-YRandom” indicates the HTTK PBPK model parameterized with a random permutation of the in vitro data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | HTPBTK-InVitro | HTPBTK-ADmet | HTPBTK-Dawson | HTPBTK-Pradeep | HTPBTK-OPERA | HTPBTK-Consensus | HTPBTK-YRandom | In Vivo Fits |
| In Vitro | 1.2 (56) | 1.3 (46) | 1.3 (44) | 1.3 (40) | 1.2 (35) | 1.1 (56) | 1.4 (56) | 0.62 (47) |
| No In Vitro |  | 1.2 (18) | 1.1 (16) | 1.1 (15) | 0.93 (22) | 0.97 (29) | 1.2 (31) | 0.55 (36) |
| QSPR Intersection | 1.3 (19) | 1.3 (25) | 1.2 (25) | 1.3 (25) | 1.2 (25) | 1.1 (25) | 1.4 (25) | 0.69 (21) |
| Maximal | 1.2 (56) | 1.2 (64) | 1.3 (60) | 1.2 (55) | 1.1 (57) | 1 (85) | 1.3 (87) | 0.59 (83) |

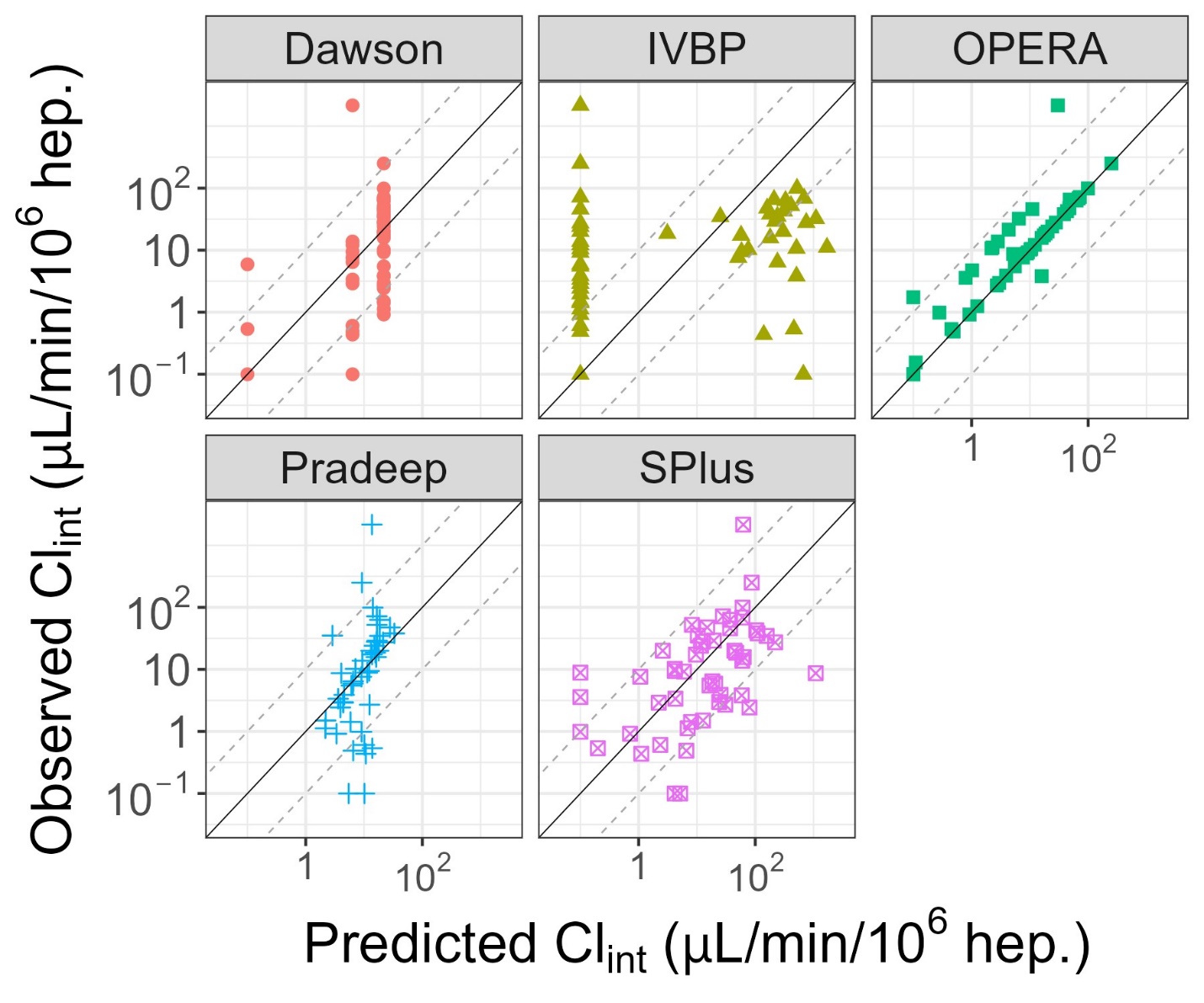
Table Level 3 Analysis of predictions based on empirical model fits and a PBPK model (“HTTK”) parameterized with chemical specific values either measured in vitro (“HTTK-InVitro”) or predicted with various QSPRs. Both the coefficient of determination (R2) and root mean squared error (RMSLE) are calculated on a log10 scale. “HTTK-YRandom” indicates the HTTK PBPK model parameterized with a random permutation of the in vitro data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Cmax** | | | **AUC** | | | | |
| **Predictor** | **R2** | **RMSLE** | **RPE** | **R2** | **RMSLE** | **RPE** | **RPE Low** | **RPE High** | |
| Empirical Fits | 0.95 | 0.95 | -0.09 | 0.96 | 0.31 | 0.32 | 0.41 | 0.06 | |
| HTTK-InVitro | 0.57 | 0.84 | -0.02 | 0.50 | 1.11 | 3.44 | 6.72 | -0.46 | |
| HTTK-ADmet | 0.62 | 0.78 | 0.15 | 0.62 | 0.97 | 5.88 | 10.50 | -0.44 | |
| HTTK-Dawson | 0.60 | 0.81 | 0.02 | 0.58 | 1.01 | 3.32 | 5.00 | -0.04 | |
| HTTK-Pradeep | 0.57 | 0.83 | -0.11 | 0.40 | 1.22 | 2.29 | 5.99 | -0.80 | |
| HTTK-OPERA | 0.58 | 0.83 | 0.11 | 0.56 | 1.04 | 4.67 | 7.93 | 0.39 | |
| HTTK-YRandom | 0.46 | 0.93 | 0.53 | 0.10 | 1.48 | 5.10 | 12.10 | -0.88 | |

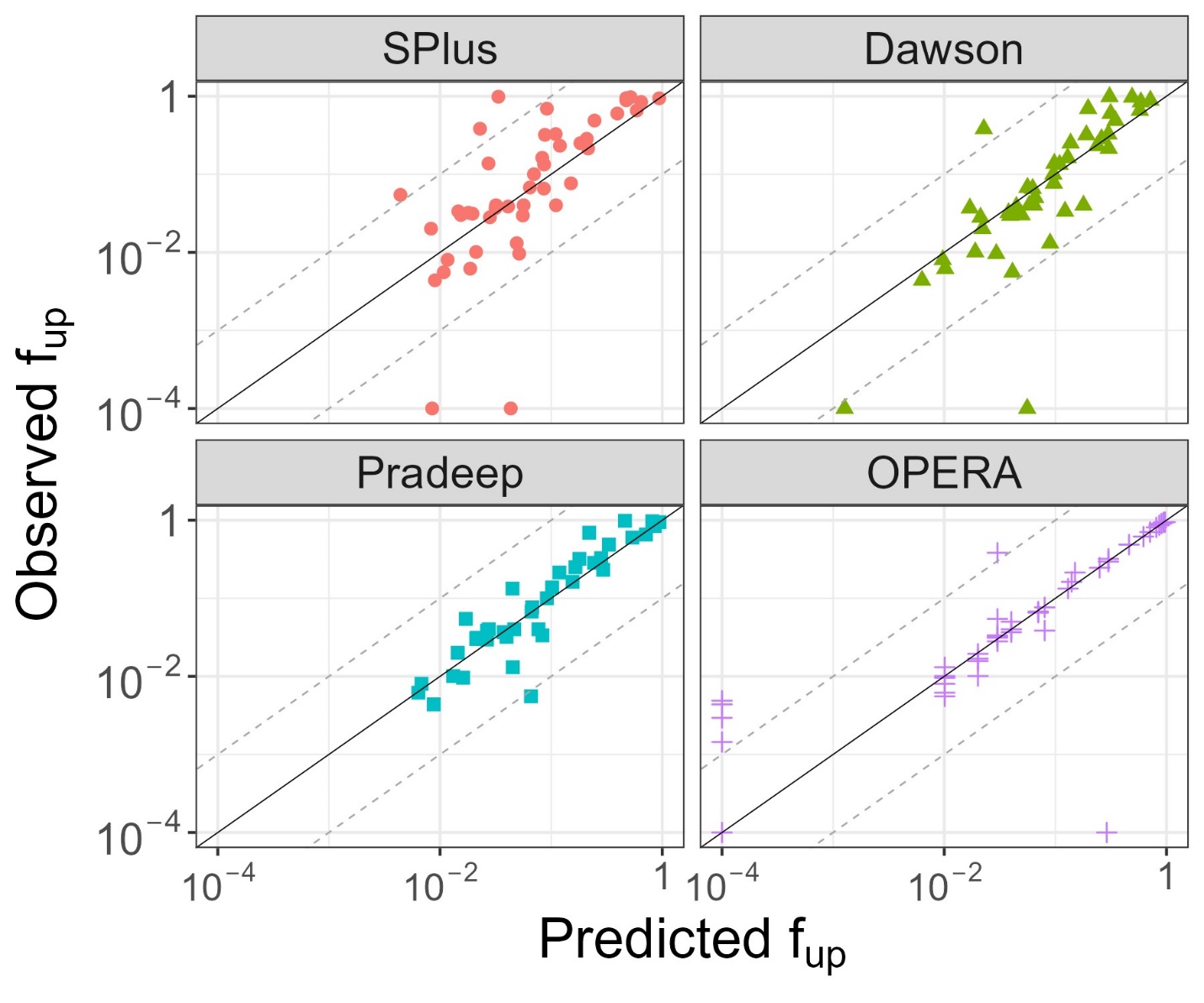
Table Analysis of predictions based on empirical model fits and a PBPK model (“HTTK”) parameterized with chemical specific values either measured in vitro (“HTTK-InVitro”) or predicted with various QSPRs. Both the coefficient of determination (R2) and root mean squared error (RMSLE) are calculated on a log10 scale. “HTTK-YRandom” indicates the HTTK PBPK model parameterized with a random permutation of the in vitro data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Level 3** | | | | | |
|  | **thalf** | | **Vd** | | **Cltot** | |
| **Predictor** | **R2** | **RMSE** | **R22** | **R222** | **R23** | **RMSE3** |
| HTTK-InVitro | 0.04 | 1.43 | 0.04 | 0.89 | 0.03 | 1.57 |
| HTTK-ADmet | 0.02 | 1.45 | 0.13 | 0.85 | 0.20 | 1.42 |
| HTTK-Dawson | 0.09 | 1.39 | 0.11 | 0.86 | 0.20 | 1.42 |
| HTTK-Pradeep | 0.00 | 1.46 | 0.13 | 0.85 | 0.02 | 1.57 |
| HTTK-OPERA | 0.09 | 1.39 | 0.05 | 0.89 | 0.09 | 1.52 |
| HTTK-YRandom | 0.01 | 1.45 | 0.00 | 0.91 | 0.02 | 1.57 |
| QSARINS | 0.10 | 1.00 | 0.04 | 0.89 | 0.19 | 1.17 |
| IFS-QAPR | 0.15 | 0.97 | 0.04 | 0.89 | 0.25 | 1.12 |

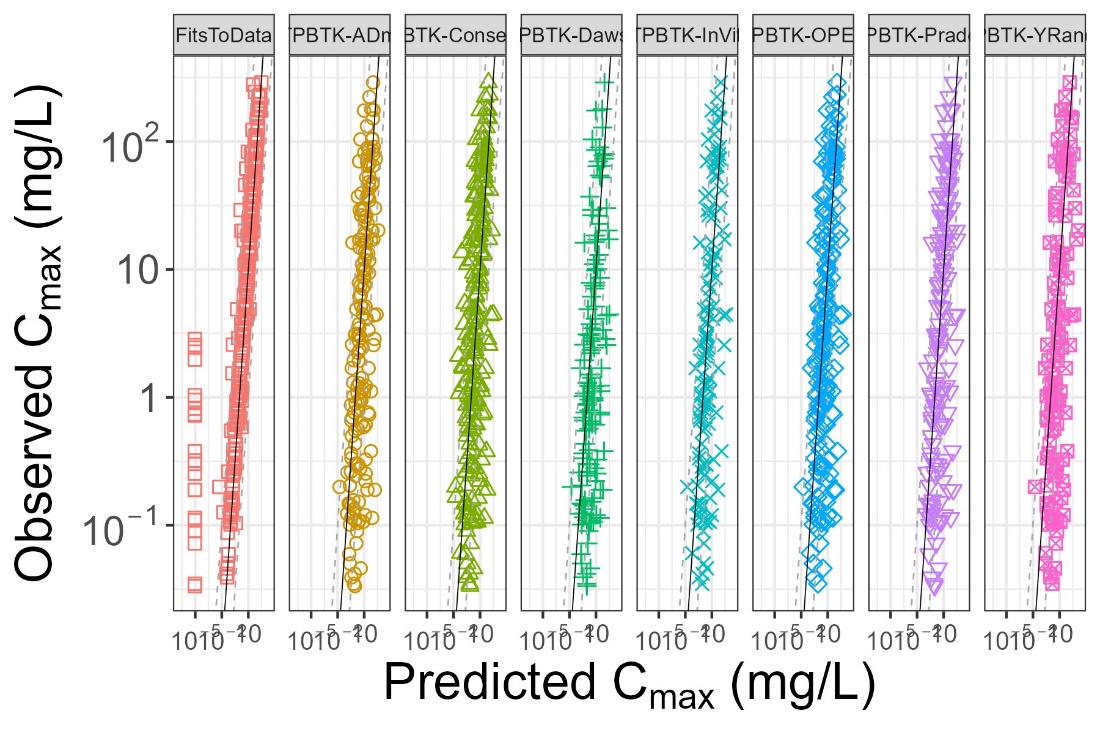
# Supplemental Figures



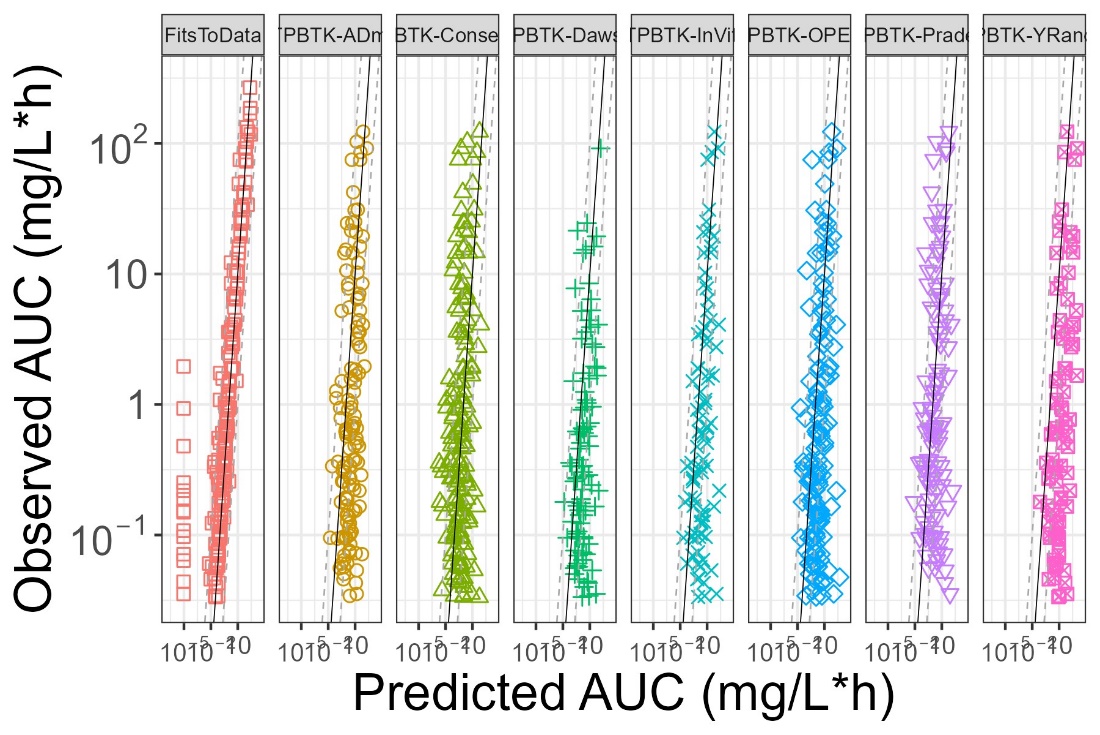
Supplemental Figure 1: Evaluation of Predictions for Intrinsic Hepatic Clearance (Clint). Zero values were plotted at 10-1, the solid line indicates identity (1:1) while the dashed lines indicate 3.2-fold difference.



Supplemental Figure 2: Evaluation of Predictions for Fraction Unbound in Plasma (fup). Zero values were plotted at 10-4, the solid line indicates identity (1:1) while the dashed lines indicate 3.2-fold difference



Supplemental Figure X: Evaluation of Predictions for Cmax based on empirical model fits (“1CompFits”), and a PBPK model (“HTTK”) parameterized with chemical specific values either measured in vitro (“HTTK-InVitro”) or predicted with various QSPRs.



Supplemental Figure X+1: Evaluation of Predictions for AUC based on empirical model fits (“1CompFits”), and a PBPK model (“HTTK”) parameterized with chemical specific values either measured in vitro (“HTTK-InVitro”) or predicted with various QSPRs.

# Supplemental Tables

Supplemental Table 1: List of 102 chemicals, physico-chemical descriptors, and *in vitro* measured values, and QSPR predictions

*SupTable-QSPRPredsandInVitroData.xlsx*

Supplemental Table 2: Concentration vs. time data (Sayre et al., 2020)

*SupTable-CvTData.xlsx*



Supplemental Table 3: Empirical (one and two compartment model) toxicokinetic parameter estimates *SupTable-TKFits.txt*



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Supplemental Table 4: Chemicals that could not be fit by either a one- or two-compartment model using R package invivoPKfitDTXSID | CAS | Model | AIC | Species | method |
| DTXSID1034187 | 10540-29-1 | model\_flat | 6.05E+08 | rat | L-BFGS-B |
| DTXSID2021315 | 1746-01-6 | model\_flat | 227 | rat | L-BFGS-B |
| DTXSID5025607 | 93-15-2 | model\_flat | 54400 | rat | L-BFGS-B |







Supplemental Table 6: Chemicals whose measured values were potentially retrieved “as is” from model training sets and were therefore removed from the evaluation: *SupTable-PossibleTrainingChems.txt*



Supplemental Table 7: Chemical-specific root mean square log10 errors for the full TK concentration time course data by QSPRs*SupTable-RMSLEbyChem.txt*

Supplemental Table 8: Level 3 Predictions

*SupTable-Level3.xlsx*