ORD MANUSCRIPT COVER SHEET

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

**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) |
| Gilberto Padilla Mercado | 1, 2 | [PadillaMercado.Gilberto@epa.gov](mailto:PadillaMercado.Gilberto@epa.gov) |
| Jon Arnot | 3 Arnot Research & Consulting | [jon@arnotresearch.com](mailto:jon@arnotresearch.com) |
| Linda Bertato | 4 University of Insubria, Varese | [l.bertato@uninsubria.it](mailto:l.bertato@uninsubria.it) |
| Trevor Brown | 3 | [trevor.n.brown@gmail.com](mailto:trevor.n.brown@gmail.com) |
| Nicola Chirico | 4 | [nicola.chirico@uninsubria.it](mailto:nicola.chirico@uninsubria.it) |
| Chris Cook | 1,2 | [cookchristopher732@gmail.com](mailto:cookchristopher732@gmail.com) |
| Daniel Dawson | 1,5 Integral | [ddawson@integral-corp.com](mailto:ddawson@integral-corp.com) |
| Sarah Davidson-Fritz | 1 | [DavidsonFritz.Sarah@epa.gov](mailto:DavidsonFritz.Sarah@epa.gov) |
| John DiBella | 6 Simulations Plus | [john.dibella@simulations-plus.com](mailto:john.dibella@simulations-plus.com) |
| Stephen Ferguson | 7 NIEHS/NTP | [stephen.ferguson@nih.gov](mailto:stephen.ferguson@nih.gov) |
| Rocky Goldsmith | 1,8 Congruence | [rgoldsmith@congruencetx.com](mailto:rgoldsmith@congruencetx.com) |
| Chris Grulke | 1 | [cmgrulke@gmail.com](mailto:cmgrulke@gmail.com) |
| Richard Judson | 1 | [Judson.Richard@epa.gov](mailto:Judson.Richard@epa.gov) |
| Michael Lawless | 6 | [mlawless@simulations-plus.com](mailto:mlawless@simulations-plus.com) |
| Kamel Mansouri | 9 Innotiv  10 NICEATM | [kamel.mansouri@nih.gov](mailto:kamel.mansouri@nih.gov) |
| Grace Patlewicz | 1 | [Patlewicz.Grace@epa.gov](mailto:Patlewicz.Grace@epa.gov) |
| Ester Papa | 4 | [ester.papa@uninsubria.it](mailto:ester.papa@uninsubria.it) |
| Prachi Pradeep | 11 German Federal Institute for Risk Assessment (BfR) | [prachipradeep.15@gmail.com](mailto:prachipradeep.15@gmail.com) |
| Alessandro Sangion | 3 | [alessandro.sangion@mail.utoronto.ca](mailto:alessandro.sangion@mail.utoronto.ca) |
| Risa Sayre | 1 | [sayre.risa@epa.gov](mailto:sayre.risa@epa.gov) |
| Russell Thomas | 1 | [Thomas.Russell@epa.gov](mailto:Thomas.Russell@epa.gov) |
| Rogelio Tornero-Velez | 1 | [tornero-velez.rogelio@epa.gov](mailto:tornero-velez.rogelio@epa.gov) |
| Barbara Wetmore | 1 | [wetmore.barbara@epa.gov](mailto:wetmore.barbara@epa.gov) |
| Michael Devito | 1 | [Devito.Michael@epa.gov](mailto:Devito.Michael@epa.gov) |

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

**Target Journal:** Leaning toward Environment International. However ALTEX is tempting.

Four bullet point summary:

* Toxicokinetic (TK) information, such as elimination half-life is critical for understanding chemical risk
* Predictions of selected *in vitro* determinants of TK were generated using several QSAR models
* The models were evaluated for their ability to reproduce *in vitro* and *in vivo* measurements of TK
* Performance of a high-throughput physiologically-based TK (PBTK) model was comparable when reliant on TK QSPRs or measured *in vitro* data

**One sentence description:** This collaborative trial demonstrates that multiple QSPRs can 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:   
Towards a consensus approach

John Wambaugh1,\*, Nisha Sipes1, Gilberto Padilla Mercado1,2, Jon Arnot3, Linda Bertato4, Trevor Brown3, Nicola Chirico4, Christopher Cook1,2, Daniel Dawson1,5, Sarah Davidson-Fritz1, John DiBella6, Stephen Ferguson7, Rocky Goldsmith1,8, Chris Grulke1, Richard Judson1, Michael Lawless6, Kamel Mansouri9,10, Grace Patlewicz1, Ester Papa4, Prachi Pradeep1,11, Alessandro Sangion3, 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, USA 28311
2. Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA 38331
3. ARC Arnot Research and Consulting Inc., Toronton, Ottawa, Canada
4. Department of Theoretical and Applied Sciences, University of Insubria, Varese, Italy
5. Integral Consulting, Seattle, Washington, USA 98104
6. Simulations Plus, Inc., Lancaster, California, USA 93534
7. Division of the National Toxicology Program (NTP), National Institute of Environmental Health Sciences, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA 28317
8. Congruence Therapeutics, Chapel Hill, North Carolina, USA 27517
9. Inotiv, Morrisville, North Carolina, USA 27560
10. NTP Interagency Center for the Evaluation of Alternative Toxicological Methods, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA 28317
11. German Federal Institute for Risk Assessment (BfR), Berlin-Jungfernheide, Germany

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

# Abstract (279 out of 250)

Chemical-specific toxicokinetic (TK) data are to assess public health risks posed by chemicals occuring in commerce and the environment*.* High throughput TK (HTTK) methods help address data gaps but typically require chemical-specific *in vitro* measurements. As an alternative to measurement, *in silico* models for HTTK have been developed. This evaluation characterizes the accuracy of HTTK approaches based on both *in vitro* measurement and *in silico* predictions. Six models from a collaborative trial were variously used to estimate intrinsic hepatic clearance (Clint), fraction of chemical unbound in plasma (fup), and/or TK elimination half-life for 88 chemicals with *in vivo* measured TK data. The Clint and fup values were used as inputs in a high throughput physiologically-based TK (PBTK) model to predict plasma concentrations. Root mean squared log10 error (RMSLE) was calculated for AUC (time-integrated area under the curve) and Cmax (peak concentration). As a point of comparison for HT-PBTK, the *in vivo* data were also used to estimate empirical one- or two-compartment TK model parameters. Early time points, including Cmax, are driven by physicochemical properties and are insensitive to HTTK parameters. Cmax could be predicted with RMSLE 0.8-0.9 in contrast to 0.7 for empirical TK models. Greater discrimination between models was observed at later time points, which impact AUC and is driven by estimated metabolism and elimination. For AUC, empirical TK models had RMSLE 0.5. Chemical-specific *in vitro* data had RMSLE 1.3 and QSPRs ranged from 1.3-1.4. A consensus prediction using the maximum Clint predicted across all QSPRs predicted AUC with RMSLE 1.1. The consensus predictions outperformed the *in vitro* measured data for the evaluation chemicals. For novel compounds, a consensus QSPR approach may yield reasonable predictions of TK.

# Introduction

Toxicokinetics (TK) describes the absorption, distribution, metabolism, and excretion (ADME) of chemicals in the body as a function of time (O'Flaherty 1981). Since TK allows the prediction of tissue concentrations as a function of chemical exposure, it provides critical information for assessing health risks posed by chemical exposures (Rotroff et al. 2010; Tonnelier et al. 2012). Whilst information on TK has been a critical aspect of human pharmaceutical safety assessment for decades, the data requirements under the majority of non-pharmaceutical regulatory frameworks do not consistently include TK or one of its underlying processes. Indeed, regulations such as the EU’s REACH have no specific information requirements to generate TK data for specific tonnage bands yet increasingly it is recognized that TK information is valuable in the interpretation of biomonitoring data (Sobus et al. 2011), dosimetric anchoring of animal toxicity studies (National Research Council 1983), quantitative *in vitro*-*in vivo* extrapolation (IVIVE) from high throughput bioactivity studies (Wetmore et al. 2015) as well as substantiating read-across justifications (Escher et al, 2019). Of the many thousands of substances that exist in commerce or in the environment, only a small percentage have been characterized in terms of their toxicity and of these, an even smaller proportion are associated with relevant TK information (Bell et al. 2018). If data does exist, or is specifically generated, animal studies have been traditionally relied upon. However, given the large number of environmental substances lacking data, the resources required to generate TK data using animals are neither practical, nor desirable from an ethics perspective especially given the call to reduce vertebrate testing under regulatory frameworks such as TSCA. Rather, new approach methodologies (NAMs) (Kavlock et al. 2018) including new experimental methods for assessing TK (Breen et al. 2021) have the potential to bridge the gap in addressing information requirements to facilitate chemical risk assessment process (National Academies of Sciences 2017).

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” (National Academies of Sciences 2017). Over the last 15 years, US Federal agencies have worked with collaborators and contractors to collect chemical-specific *in vitro* data that could permit the estimation of TK parameters (Paini et al. 2020; Rotroff et al. 2010; Tonnelier et al. 2012; Wambaugh et al. 2019; Wetmore et al. 2015; Wetmore et al. 2013; Wetmore et al. 2012). *In vitro* bioactivity data and points of departure are now available for thousands of chemicals notably arising from the ToxCast (Richard et al. 2016) and Tox21 (Tice et al. 2013) screening programs. Experimental *in vitro* TK data has been generated for approximately 1000 of the ToxCast and Tox21 chemicals permitting calculations by IVIVE (Bell et al. 2018; Breen et al. 2021; Coecke et al. 2013; Wetmore 2015) to be made to determine what concentrations in tissues would be expected based on the active *in vitro* concentrations. Although these *in vitro* TK approaches have significantly enhanced the throughput and decreased both cost and animal usage, the scale of the number of chemicals without such TK information is still not viable from a cost and testing perspective. These challenges motivated this study to investigate to what extent *in silico* models could be used to predict key TK parameters *in vitro* which could then be used in *in vivo* PBTK models. This is synergistic to how the pharmaceutical industry has moved away from animal studies by relying on characterizing certain aspects of TK *in vitro* and then extrapolating those to *in vivo* conditions to estimate key TK parameters such as AUC (Wang 2010).

Multiple *in silico* models specifically quantitative structure-property relationship (QSPR) models have been developed and published that predict TK parameters such as fraction unbound and inherent clearance (Chirico et al. 2021; Dawson et al. 2021; Kirman et al. 2015; Pradeep et al. 2020; Sipes et al. 2017). The QSPRs span the range of open-sourced models to freely available or being commercial and underpinned by pharmaceutical and non-pharmaceutical chemistries. Each model has its own training and testing datasets comprising a range of chemicals with different physicochemical properties. Performance metrics reported differ across models available. These factors make comparing and contrasting the scope and utility of these different models to make reliable and robust TK parameter estimates a challenge. There is a need to assess the performance and coverage of these individual QSPR models in predicting TK parameters for chemicals occuring in commerce and the environment.

Here we have evaluated the TK parameter estimates originating from models developed by six modeling groups. Starting from an initial list of researchers interested in HTTK who participated in the ExpoCast community of practice (Wambaugh and Rager 2022) and Tox21 (Tice et al. 2013), a presentation was made to 19th International Workshop on (Q)SAR in Environmental and Health Sciences in June, 2021 which further solicited involvement from the HTTK QSAR community. Participants represent different international academic, regulatory, and commercial entities. Four modeling groups produced QSPR models to predict intrinsic hepatic clearance (Clint measured with hepatocyte incubations, (Shibata et al. 2002)) and fraction unbound in plasma (fup, typically measured via rapid equilibrium dialysis (Waters et al. 2008)). Both parameters are important in PBTK analyses for generating AUC and Cmax. In this study, the models developed were initially evaluated for their ability to reproduce *in vitro* measured values. Then an analysis of QSPR-based predictions for chemical concentration as a function of time (CvT) was undertaken. A generic physiologically-based TK (PBTK) model (Pearce et al. 2017b) was used with the predictions from each QSPR. *In vivo* plasma and blood CvT data for rat and human for 101 chemicals from the CvTdb (Sayre et al. 2020) formed the basis of the dataset used. Models were evaluated for their ability to reproduce the full CvT curve, summary statistics (such as AUC and Cmax) as well as TK parameters such as chemical half-life. Two additional models for chemical half-life were also evaluated. The six modeling groups were provided with the chemical identities (including their chemical structures) and selected physico-chemical descriptors (SMILES structure descriptors (Weininger 1988) and OPERA physico-chemical predictions (Mansouri et al. 2018)) but were not provided with the actual *in vivo* evaluation data.

# Methods

Table 1 summarizes the data used for evaluation and the statistical approaches used, while Table 2 describes the human HTTK parameter QSPRS that were evaluated. Evaluation data were screened for suitability based on their ability to be described by empirical (that is, compartmental) TK models. QSPR evaluation was then conducted at four levels: 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. At (Level 3) the ability of the QSPRs and PBTK model to predict summary statistics (for example Cmax, AUC, half-life) was evaluated. Finally (Level 4) the impact of error in HTTK-based predictions was evaluated by systematically substituting HTTK parameter predictions for parameters optimized to the *in vivo* data. All analyses were performed in the free, open-source statistical analysis language R (R Core Team 2021) v4.5.0 (the current developer version). Heatmaps were generated using ggplots::heatmap.2 (Warnes et al. 2024) with data clustered on the basis of Euclidean distance via base R function stats::dist (R Core Team 2021) . All code is documented using RMarkdown (Baumer and Udwin 2015) and available as supplemental material at <https://github.com/USEPA/CompTox-ExpoCast-HTTKQSPRs>.

## Evaluation Data and Analysis

Table 1 Chemical Counts and Evaluations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step | Description | Chemicals for Evaluation | QSPR Models Evaluated | Reference |
| Initial Chemical List | Chemicals with Cvt data following oral and/or intravenous dosing. Modelers were provided with chemical identity and physico-chemical properties. | 101 |  | (Sayre et al. 2020) |
| Pre-Screen for Empirical TK Model Parameter Estimation | Chemical data sets were eliminated for having too few points above the limit of detection | 95 |  | (Mercado et al. 2024) |
| Data Filtering Based on Empirical Parameter Estimations | Chemicals that could not be described with a one- or two-compartment empirical PK model were withheld from further analysis. | 92 with good empirical model fits |  | (Mercado et al. 2024) |
| Level 1 | Direct evaluation of HTTK QSPR predictions using *in vitro* TK Measurements (fup, Clint) | 64 with Measured *In vitro* Data | 4 (fup),  5 (Clint) | (Pearce et al. 2017b) |
| Level 2 | Evaluation of prediction for full TK time course using HT-PBTK with HTTK QSPR (all time points) | 17-88 depending on chemicals compared and number of available predictions | 5 + Consensus | (Sayre et al. 2020) |
| Level 3 | Evaluation of predicted TK summary statistics  (Cmax, time-integral/AUC, Vd, thalf, Cltot, Css) | 17-88 | 6 + Consensus | (Sayre et al. 2020) |
| Level 4 | Sensitivity of error to substituting empirically-estimated values for absorption, distribution, or metabolism/elimination with values from consensus model | 88 | Consensus |  |
| Level 2 and 3 In Vitro-Parameterized HT-PBTK | Chemicals for which *in vitro*-measured Clint and Fup were both available were also evaluated as a point of reference | 58 |  | (Pearce et al. 2017b) |
| Interspecies Evaluation | Chemicals with Clint and Fup measured in both rat and human were used to determine the impact of using human QSPRs to make predictions in rat | 130 |  | (Honda et al. 2019; Wetmore et al. 2013) |

### Initial Chemical List

Varying amounts of chemical-specific data were available for the different evaluations performed. Table 1 describes the number of chemicals and type of evaluation that was performed at each. At the start of in the public release of CvTdb re available for 101 chemicals from EPA’s CvT database (Sayre et al. 2020). To be included either rat or human time course *in vivo* blood or plasma concentration data following oral and/or intravenous doses were required. The *in vivo* measured time course values are available as Supplemental Table 2. As described late in the Methods, chemical- and species-specific parameters were estimated for empirical (compartmental) TK models. As described in Results, chemical-species combinations that could not be adequately described by simple TK models were excluded from subsequent analysis.

### Data Filtering Based on Suitability for Modeling

The CvTdb aims to fully document TK concentration vs. time experiments in the public domain. Not all these experimental data sets are suitable for TK modeling. TK concentration data may be confounded by limit of detection (LOD) and limit of quantitation (LOQ). LOD refers to the minimal signal the chemical analysis method can detect, while LOQ refers to the minimal signal that can be interpreted quantitatively as a chemical concentration. In some cases, a chemical-species combination has too few observations above the LOQ to allow meaningful modeling. These datasets were filtered out by invivoPKfit prior to attempting parameter estimation (Mercado et al. 2024).

### Data Filtering Based on Empirical Parameter Estimates

Parameters were estimated for empirical one- and two-compartment toxicokinetic models for chemicals with available CvT data using R package “invivoPKfit” (<https://github.com/USEPA/CompTox-ExpoCast-invivoPKfit>) (Mercado et al. 2024). 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 (Akaike 1998). 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 parameterized by either *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.

### Level 1 Analysis

Chemicals for which *in vitro* TK Measurements (fup, Clint) were available allowed direct evaluation of HTTK QSPR predictions. These data are drawn from the peer reviewed scientific literature and included in the R package “httk” by default. It is possible 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. Kolmogorov-Smirnov tests for the Level 1 analysis were performed using R function *ks.test*.

### Level 2 Analysis

The chemical-specific HTTK QSPR parameters were incorporated into a generic “high throughput” PBTK (or HT-PBTK) model designed for use with HTTK parameters. The HT-PBTK model predicted concentration vs time outcomes in plasma (and other tissues, although the evaluation data were all plasma) as a function of species, dose route, dose amount, and observation times. The physiological aspects of the model were varied based on species specific scaling estimates (Davies and Morris 1993) in units of L/h/kg3/4 for flows and L/kg for volumes. 70 kg was used for the human body weight and 0.25 kg was used for the rat.

R package “httk” (Pearce et al. 2017b) v2.3.2 was used to provide HT-PBTK predictions for the Level 2 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). A generic PBTK model was used (model “gas\_pbtk”) that consists of well-mixed compartments for the gut, kidney, liver, lung, and rest of body. The model allows exposure by oral and intravenous doses, and clearance by glomerular filtration (kidneys), metabolism (liver), and exhalation (lungs). Though the model allows for inhalation exposures, these were not present in the CvT data used. The model is parameterized for a chemical using fup and Clint plus equilibrium tissue:plasma partition coefficients predicted with a modified Schmitt’s method (Pearce et al. 2017a; Schmitt 2008) including aspects of Peyret et al. (2010). Clintin vitro(units of µL/min/106 hepatocytes) is scaled to Clintliver (L/h/kg) using species specific values for liver volume, density, and hepatocellularity. Oral dosing is subject to first-pass metabolism by the liver before the compound distributes systemically. 100% oral absorption from the gut was assumed, with no gut metabolism. The model was simulated using command “httk::solve\_gas\_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.

### Level 3 Analysis

This level of analysis focuses on the impact of using HTTK QSPRs to predict TK summary statistics – key aspects of the CvT curve that often characterize the implications of TK. The predictions were evaluated by using values estimated from the *in vivo*-calibrated empirical models. These statistics include peak concentration (Cmax), time-integrated concentration (AUC), steady-state concentration resulting from repeated dosing (Css), total clearance (CLtot), elimination half-life (thalf), and volume of distribution (Vd). Based on QSPR-derived and *in vitro* parameters, these summary statistics were calculated using various functions of R package “httk”: Volume of distribution was calculated using the same calibrated Schmitt partition coefficient algorithm as for the HT-PBTK model, but with all tissues lumped into a single compartment. Volume of distribution depends largely on physico-chemical properties with some input from fup and was calculated using httk::calc\_vdist(). For kelim the inverse of the steady-state concentration (1/Css) was used as an effective clearance rate and was converted to elimination rate using the estimated (Vd). Steady-state concentration depends upon both Clint and fup and was calculated using httk::calc\_css(model=”gas\_pbtk”) so as to include exhalation as a route of elimination for semi- and volatile chemicals.

Human half-life predictions (IFS-QSAR and QSARINS-Chem) were first converted to rat predictions using allometric scaling (the ratio of rat body weight, assumed to be 0.25 kg, to human, 70 kg, raised to the ¾ power). Half-lives were converted to elimination rates and multiplied by the Vd from the consensus model to estimate total clearance and Css.

### Level 4 Analysis

Level 4 is an uncertainty analysis. We identified the relative contribution of HTTK errors in different phases of TK (Figure 1) to the overall error. Empirical one- or two-compartment models were used (as indicated by invivoPKfit). The source of the parameters for the compartmental models was varied between in vivo-data based and values predicted by HT-PBTK using the consensus QSPR predictions. Level 4 focused only on the consensus model values for HTTK parameters. The empirical TK model parameters estimated directly from the in vivo data were assumed to represent the lowest overall error. Sensitivity of error was calculated by sequentially substituting values derived from the consensus HTTK model for the in vivo-estimated values for absorption (Fbio), distribution (Vd), or metabolism/elimination (kelim). Calculation of Fbio was assumed to only include first-pass hepatic metabolism. First pass metabolism for Fbio depends on Clint and fup, and was calculated using httk::calc\_hep\_bioavailability().

### Evaluation Metrics

At each level 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 = .

## QSPR Models Evaluated

Table 2 QSPR Models Evaluated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Predictions | Original Units | Mechanism | Reference |
| Simulations Plus ADMET Predictor® | Level 1  (*in vitro* parameters) | μL/min/mg of microsomal protein | Sum of CYP-specific Artificial Neural Network (ANN) | (Sipes et al. 2017) |
| Pradeep 2020 | Level 1 | µL/min/106 hepatocytes | Random forest and support vectors method | (Pradeep et al. 2020) |
| Dawson 2021 | Level 1 | µL/min/106 hepatocytes | Random forest, clearance organized by categories | (Dawson et al. 2021) |
| OPERA | Level 1 | µL/min/106 hepatocytes | Nearest-neighbors? | (Mansouri et al. 2021; Mansouri et al. 2018) |
| IFS-QSAR | Level 3  (Half-lives) | h | Fragment-based Multiple Linear Regressors (MRL) | (Arnot et al. 2014) |
| QSARINS-Chem | Level 3  (Half-lives) | h | Ordinary Least Squares MLR | (Papa et al. 2018) |
| *In vitro* Biotransformation Prediction-Suite (IVBP-Suite) IVBP | Level 1 (Clint only) | log10 mL/h/10^6hep | IVBP predictions are based on SMARTCyp (Rydberg et al. 2010) ranks which were tested for CYP3A4, CYP1A2, CYP2C19 substrates, and they should work also for isoforms 2A6, 2B6, 2C8. | (Chirico et al. 2021) |

The QSPR models evaluated are summarized in Table 2. Four different modeling groups previously (Dawson et al. 2021; Pradeep et al. 2020; Sipes et al. 2017) 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). 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, predictions that seemed to be a direct retrieval of the chemical-specific values from a training set were removed from the evaluations. 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.. Omitted predictions are listed in Supplemental Table 6.

QSPR-specific predictions were made by altering the values used by httk. httk draws values from a asingle table which stores the fup and Clint values for all chemicals (httk::chem.phys\_and\_invitro.data). After alteration, the httk functions 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)” (Dawson et al. 2021; Pradeep et al. 2020; Sipes et al. 2017). 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.

## Consensus Model

A consensus QSPR was constructed for fup by taking the inverse logit of the mean of the logit-transformed fup from each QSPR. A consensus QSPR was constructed for Clint by using the maximum Clint value predicted by any QSPR.

## Interspecies Concordance

The evaluation data in Table 1 are largely drawn from *in vivo* experiments with rat, while the HTTK parameter QPSRs in Table 2 are trained to predict values for human. To investigate the impact of the mismatch between HTTK and in vivo data, a set of chemicals with in vitro HTTK measurements for both human and rat were investigated. That data, reported by R package httk, were largely drawn from (Wetmore et al. 2013) and (Honda et al. 2019). First, the error in using human-measured Clint and fup to predict rat-measured Clint and fup was characterized. Then the two sets of measurements were used to predict Cmax and Css, in both cases using the same rat-specific physiological parameters and scaling described above. This allowed estimation of the error introduced by using human-derived HTTK measurements to make rat-specific HTTK predictions.

# Results

## Evaluation Chemicals and Predictions

There are 101 chemicals present in the public release of CvTdb (Sayre, 2020) that had chemical concentration vs. time (CvT) data resulting from either oral or intravenous doses given to rats or humans. These chemicals included: 57 from the non-confidential Toxic Substances Control Act (TSCA) active inventory (Schmidt 2016), 20 pharmaceuticals, 24 pesticides, 99 that are found in consumer products, 7 per- and poly-fluoroalkyl substances (PFAS) (Patlewicz et al. 2019), and 64 that are part of the ToxCast screening program. Note that a chemical could be in more than one of these categories.

As illustrated in Figure 1, a typical TK concentration vs. time curve following a single oral dose consists of separate phases: 1) an initial increase driven by absorption of the chemical into the body, 2) a steep decrease driven by distribution of the chemical from the blood into tissues with varying affinities for the chemical, followed by 3) metabolism and excretion of the chemical (elimination) from the blood. Metabolism and excretion are present during the distribution phase. The distribution phase can be pronounced depending on how much chemical the tissues can sequester. Following an intravenous dose there is no absorption phase, although there may be a very short initial phase as the chemical is distributed from the site of injection to the systemic blood circulation. The different phases of TK vary in sensitivity to the HTTK parameters analyzed here. The elimination phase is most sensitive because HTTK parameters allow prediction of glomerular filtration in the kidney (excretion) and metabolic clearance in the liver (metabolism).

CvT data sets were modeled for their suitability to evaluate HTTK predictions by systematically optimizing parameters for one- and two-compartment models (Mercado et al. 2024). That is, we assumed data sets that could not be described by a one- or two-compartment model did not conform to the general profile depicted in Figure 1. CvT data sets for six chemicals were eliminated in pre-processing before fitting: C.I. Solvent Red 1 and Propylparaben were eliminated because there were no observations above LOQ. The data for Phenazone (antipyrine in CvTdb) are from radiolabeled experiments which only allow estimation of elimination rate (that is, no concentration data are available). Tamoxifen has no available LOQs and the majority of observations are non-detects. Pentachloroanisol has only two timepoints. 2,3,4,7,8-Pentachlorodibenzofuran has all measured concentration values below the limit of detection.

Of the 95 chemicals where empirical TK model parameters were estimated, data for 3 chemicals were best described by the “flat” model indicating that the data sets were too noisy to estimate empirical TK parameters: Perfluorohexanoic acid, Perfluorooctanesulfonate, and Perfluorooctanoic acid Given that these chemicals were poorly described by basic TK models these chemicals were withheld from subsequent analysis. It is likely that sex-specific differences, which are currently not accounted for the parameter optimization software used (Mercado et al. 2024) but are known to be influential in PFAS (Fenton et al.) caused the problems for these three chemicals.

The data for each remaining chemical could be described using either a one- or two-compartment empirical models. There were 11 chemicals with only oral data. For these oral-only datasets we could not estimate all the parameters needed to make CvT predictions. Separate parameter estimates were made for each combination of compound and species for which there were data. For each chemical, the one or two-compartment models was selected based on AIC.

## QSPR Predictions

The QSPRs evaluated are described in Table 2. The number of chemicals for which predictions 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. For example, only a subset of chemicals had predictions from every QSPR – we denote this subset as “intersection”. Additionally, measured HTTK *in vitro* data were only available for some chemicals – we denote this subset as “in vitro”. Accordingly, evaluations aimed at characterizing the predictivity of *in vitro* HTTK itself are reported only for the “in vitro” subset. Finally, statistics are reported for the maximal number of chemicals for which the QSPR could make predictions – denoted as “maximum”. The intersectional subset of chemicals with overlap between all models is more rigorous statistically across QSPRs. The maximal superset of all chemicals with predictions is more rigorous chemically for each QSPR.

For each QSPR we removed predictions where the predicted values for a given chemical were within 1% for both measured fup and Clint. We assuming these removed predictions reflected the chemical data present in the training set and that the QSPR methodology allowed for direct recall of the measurements. At total of 21 predictions, all made by OPERA, were removed for possibly being in the training set (see Supplement Materials Table SupTable-PossibleTrainingChems.txt).

We summarize the chemical-specific properties and predictions in Figure 2, where similar chemicals (rows) and properties/predictions (columns) are clustered together based on Euclidean distance. 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.

## Level 1 Analysis

Our first level of evaluation directly compared the predictions of QSPR's with the *in vitro* measured values. Since HTTK data for many of these chemicals are present in the training sets for the QSPRs, this is not an external evaluation testing the performance of the QSPRs. However, this Level 1 analysis does allow characterization of the differences between the models.

Supplemental Table SupTable-QSPRPredCounts.txt gives the per QSPR number of chemicals for which predictions were made. For Clint the number of chemicals with HTTK parameter predictions ranged from a low of 53 for the Pradeep et al. (2020) QSPR to 88 for OPERA. For fup Pradeep et al. (2020) predicted a low of 53 and ADMET covered a high of 85. Evaluation data were available for Clint for 64 chemicals and fup for 59. Table 3 summarizes the fold errors by parameter and QSPR. Note that there were 21 potential training chemicals that were removed from all analyses because they had been predicted too accurately.

Table 3: 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 | 50 | 0.575 | 0.0239 | 45 | 0.287 | -0.116 |
| Dawson | 48 | 0.362 | -0.0245 | 43 | 0.15 | 0.0115 |
| Pradeep | 43 | 0.232 | 0.00985 | 38 | 0.137 | -0.0429 |
| OPERA | 43 | 0.00117 | -0.00014 | 38 | 0.0296 | 1.09E-05 |
| IVBP | 51 | 0.384 | 0.249 | 0 |  |  |

We evaluate model performance for Clint in Figure 3. The QSPR models perform reasonably similarly. Note that the Dawson (2021) model is categorical (that is, predicting only three values: very slow, slow, and fast) while the other models are continuous. Inclusion of non-pharmaceutical chemical data from the ToxCast project decreased QSPR error for the evaluation chemicals. The median 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. The QSPRs trained more broadly, Simulations Plus ADMET predictor and IVBP, had larger median RSMLEs. In Table 3 we summarize the fold errors for the five QSPRs. 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.

Figure 4 shows that all four models perform very well for predicting fup. Predictions are highly correlated with observations.

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 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.

## Consensus Model

Due to the varying domains of applicability of the QSPR models, there were only 23 chemicals in the intersection subset where all QSPRs could make a prediction. However, there were 88 chemicals where at least one of the QSPRs could make a prediction. A consensus model was constructed using the predictions of all QSPRs. When multiple QSPR predictions were available they were combined into a single value. For plasma protein binding the predictions were averaged, however for hepatic clearance the highest value was used. For clearance the thought was that if any one model predicted rapid metabolism that the compound should be assumed to be rapidly metabolized.

## Level 2 Analysis

For the second level of analysis, we compared predictions based on the QSPR predicted values with actual CvT data. We show all CvT curve fits and predictions on a per QSPR, chemical, species, and route basis in the Supplemental materials (Supplemental Figures SupFig-ChembyQSPRCvTPlots.pdf).

All QSPR predicted values were used with a HT-PBTK model to make CvT predictions. We bracket QSPR model performance in three ways: First, we use the HT-PBTK model with the actual *in vitro* measured values. In subsequent figures this is labeled as “HTTK-InVitro”. While there were 64 chemicals with measured Clint and 59 chemicals with measured fup, for a total of 65 unique chemicals, there were only 58 with both measured values available. *In vitro*-parameterized HT-PBTK was only evaluated for this set of 58 chemicals. Next, for a best-case performance, we use empirical (one- or two-compartment TK model) predictions based on parameterized optimized (or “fit”) 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”. We do not y-randomize the physico-chemical properties so as only to examine the parameters being predicted by the HTTK QSPRs in Table 2.

Figure 5 shows the full predicted time-courses for each set of QSPR 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 Figure 5, indicating that the predicted concentrations were relatively constant (and low) over time while the observed concentrations changed. Typically, these are chemicals where the CvT time course was especially biphasic, with an initial rapid decline and then long tails where low levels of the chemical remained for a long time. In the tails the models tend to underestimate concentration, even for the *in vivo* fits.

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 reported in Table 4. Optimal performance is given by the empirical fits to the data. Worst case performance is given by the *y*-randomized measured data. 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.

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.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | HT-PBTK-InVitro | HT-PBTK-ADMET | HT-PBTK-Dawson | HT-PBTK-Pradeep | HT-PBTK-OPERA | HT-PBTK-Consensus | HT-PBTK-YRandom | *In vivo* Fits |
| *In vitro* | 1.2 (58) | 1.3 (45) | 1.3 (43) | 1.3 (38) | 1.2 (34) | 1.1 (58) | 1.4 (58) | 0.65 (50) |
| No *In vitro* | NA (0) | 1.2 (18) | 1.1 (15) | 1.1 (15) | 0.84 (23) | 0.93 (30) | 1.2 (33) | 0.73 (30) |
| QSPR Intersection | 1.4 (17) | 1.4 (23) | 1.3 (23) | 1.3 (23) | 1.3 (23) | 1.2 (23) | 1.4 (23) | 0.58 (20) |
| Maximal | 1.2 (58) | 1.3 (63) | 1.3 (58) | 1.3 (53) | 1 (57) | 1.1 (88) | 1.3 (91) | 0.68 (80) |

The QSPRs perform roughly equally, with RMSLE generally ranging between 1.1 and 1.3 in Table 4. Absolute Average Fold Error results were also calculated and are summarized in Supplemental Table SupTable-Level2-Cvt-AAFEstats.txt. 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.

In Table 4 we note that error was roughly consistent regardless of 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.

In Figure 6 we examine the distribution of per chemical RMSLE. In the first panel of Figure 6 all observed time points are valued equally, without consideration of phase (absorption/distribution/metabolism) and measurement accuracy. In Figure 6 the performance of HT-PBTK with parameters for a random chemical, while worse, is not a marked departure from the performance HT-PBTK with the correct parameters. At early time points (second panel of Figure 6) all methods are more accurate than for all time points (first panel). This is consistent with Figure 1: The early absorption and distribution phases are dominated by the volume of distribution (Vd). Prediction of Vd largely depends 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 based on the mean absorption observed across the chemicals profiled in Wambaugh et al. (2018).

Also consistent with Figure 1, the most discriminating data for judging HTTK-based CvT predictions are the later time points which characterize metabolism and excretion (the elimination phase of TK). In the third panel of Figure 6 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 than the QSPRs. The HT-PBTK predictions based upon *in vitro*-measured HTTK data have an RMSLE which is indistinct from the QSPRs. Remarkably, the consensus QSPR predictions, using the most rapid predicted clearance, outperform the *in vitro* data, with an RMSLE of 1.1 across the evaluation chemicals.

Figure 7 shows the chemical-specific RMSLEs as a function of prediction method. In Figure 7 each column indicates a different chemical and each row a different method. These data are also provided in Supplemental Table SupTable-RMSLEbyChem.txt. Chemicals and methods have been clustered based upon Euclidean distance. White gaps mark chemicals that were not in the domain of applicability of different QSPRs (or, in the case of *in vitro* data, no measurements were available). We see that the consensus QSPR also gives the best coverage of chemicals, because the domain of applicability of the different QSPRs are varied, but sometimes it has the worst RMSLE for certain chemicals. In general, the RMSLE for a given chemical appears to be similar across model predictions; however, there are several chemicals where the *in vivo* fits are substantially better.

## Level 3 Analysis

The third level evaluates prediction of TK summary parameters; specifically, peak concentration (Cmax), area under the plasma concentration time course (AUC), Vd, half-life for elimination from the body (thalf), and whole-body clearance (Cltot). Where available, we 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 SupTable-Level3.txt.

Early time points are dominated by the ability to correctly predict peak plasma concentration (Cmax). In Table 5 we examine each methods accuracy in predicting Cmax as determined from the CvT data. Empirical fits are again best. Predictions based on *in vitro* measured HTTK data are not that different from QSPR predictions. y-randomization shows that Cmax is relatively insensitive to fup and Clint, with RMSLE of 1. That is, we do not see large differences between *in vivo* data, QSPRs, or y-randomization with respect to Cmax. As shown in Supplemental Figure 3, all the models tend to do a good job predicting Cmax greater than 1 mg/L but have a tendency to overestimate when Cmax is less than 1 mg/L. Cmax depends on Vd (for intravenous doses Cmax = dose / Vd). Given that the CvT data were relatively insensitive to HTTK parameters at early time points (Figure 6), as shown in Table 6, y-randomization was not much worse for predicting Vd when compared with both *in vitro* data and QSPR-based predictions.

Table 5 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.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | HT-PBTK-InVitro | HT-PBTK-ADMET | HT-PBTK-Dawson | HT-PBTK-Pradeep | HT-PBTK-OPERA | HT-PBTK-Consensus | HT-PBTK-YRandom | In vivo Fits |
| Cmax RMSLE | 0.82 (58) | 0.92 (45) | 0.99 (43) | 0.98 (38) | 0.77 (34) | 0.81 (58) | 1 (58) | 0.7 (50) |
| AUC RMSLE | 1.3 (58) | 1.4 (45) | 1.5 (43) | 1.5 (38) | 1.3 (34) | 1.1 (58) | 1.9 (58) | 0.62 (50) |

The most discriminating data for judging HTTK-based CvT predictions depend on the later time points which characterize metabolism and excretion and inform metrics such as AUC. In Table 5 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. *In vitro* measured HTTK data predict AUC with an RMSLE of 1.3, while the QSPRs range from RMSLE 1.3 to 1.5. The consensus QSPR predictions were the best model, with an RMSLE of 1.1.

Table 6 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. RMSLE is calculated on a log10 scale, such that RMSLE = 1 indicates ten-fold error, RMSLE = 2 indicates one hundred-fold error.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | HT-PBTK-InVitro | HT-PBTK-ADmet | HT-PBTK-Dawson | HT-PBTK-Pradeep | HT-PBTK-OPERA | HT-PBTK-Consensus | HT-PBTK-YRandom | QSARINS | IFS-QSAR |
| Half-Life RMSLE (In Vitro HTTK) | 1.6 | 1.5 | 1.5 | 1.5 | 1.3 | 1.7 | 2.7 | 0.71 | 0.83 |
| Half-Life RMSLE (QSPR Intersection) | 1.3 | 1.2 | 1.2 | 1.3 | 1.4 | 1.3 | 2.4 | 0.93 | 0.96 |
| Cltot RMSLE (In Vitro HTTK) | 1.7 | 1.6 | 1.6 | 1.5 | 1.7 | 1.5 | 2.5 | 1.3 | 1.1 |
| CLtot RMSLE (QSPR Intersection) | 1.4 | 1.3 | 1.3 | 1.3 | 1.5 | 0.92 | 2.4 | 1 | 1.1 |
| Vd RMSLE (In Vitro HTTK) | 1.4 | 1.4 | 1.4 | 1.2 | 1.3 | 1.4 | 1.1 |  |  |
| Vd RMSLE (QSPR Intersection) | 1 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.1 |  |  |
| Css RMSLE (In Vitro HTTK) | 1.4 | 1.3 | 1.3 | 1.3 | 1.3 | 1.5 | 1.9 | 1.3 | 1.1 |
| Css RMSLE (QSPR Intersection) | 1.4 | 1.3 | 1.3 | 1.3 | 1.5 | 0.92 | 2.4 | 1 | 1.1 |

In Table 6, we also examine two quantities that inform our ability to predict AUC at late time points -- thalf and CLtot examine predicted vs. observed thalf. Most of the models and the *in vitro* measured data were unsuccessful for predicting thalf, with RMSLE indicating errors between 1.3 and 1.6 (20-fold to 40-fold). The QSPR models that were trained specifically to thalf, data (rather than *in vitro* measured Clint and fup) produce notably better predictions, with RMSLE indicating errors less than a factor of ten. Finally, in Table 6 we examine predictions for CLtot, which depends on both elimination rate (inverse of thalf) and Vd. The consensus model and ADMET had the smallest errors for Css and Cltot (which are related quantities).

## Level 4 Analysis

Table 6 indicates that different aspects of TK (as indicated by Figure 1) are predicted with varying accuracy by QSPR-parameterized HT-PBTK. To evaluate the impact of the errors summarized in Table 6, we used the parameters estimated directly from the *in vivo* data as the “best” case and then substituted QSPR-based values for different parameters. In Figure 8 we first indicate the distribution of per-chemical RSMLE when a one- or two-compartment model is used with parameters optimized to the *in vivo* data. We then substitute parameters based on the consensus QSPR and HT-PBTK for the absorption (Fbio), distribution (Vdist), and elimination (kelim) phases of TK. We see that the largest errors are observed when we use HTTK for the elimination rate, consistent with the idea that estimation of metabolism is the most challenging aspect of HTTK. In Figure 8 we also include the distribution of RMSLE when all parameters are derived from the consensus QSPR – the errors from each phase of TK combine to produce a larger error, although the median error is still within a factor of ten.

## Interspecies Concordance

The evaluation data in Table 1 are largely drawn from rat *in vivo* experiments, while the HTTK parameter QPSRs in Table 2 are trained to predict values for human. The impact of the mismatch between HTTK and *in vivo* data was characterized using a set of chemicals with in vitro HTTK measurements for both human and rat. There were 130 chemicals with *in vitro* measured fup and Clint for both species. These HTTK data, reported by R package httk, were mostly drawn from Wetmore et al. (2013) and Honda et al. (2019). There were no *in vivo* data used in this interspecies analysis. This analysis allowed estimation of the error introduced by using human-derived HTTK measurements to make rat-specific HTTK predictions.

First, the error in using human-measured Clint and fup to predict rat-measured Clint and fup was characterized. As shown in Figure 9, *in vitro* measured values were roughly concordant, with RMSE (arithmetic scale) of 0.145 for fup and 0.957 for Clint. Of note in panel b is that there are chemicals with observed clearance in one species and not the other – values of Clint = 0 were plotted at 10-3 to be visible on the logarithmic scale.

Then the two sets of measurements were used to predict Cmax and Css, in both cases using the same rat-specific physiological parameters and scaling as with the Level 3 analysis. HT-PBTK-based TK predictions indicated relatively small errors due to interspecies differences in Clint and fup. As shown in Figure 9 panels c and d, RMSLE for Css was 0.307. RMSLE for Cmax was 0.239. These errors directly indicate that the errors reported in Table 5 may be slightly overestimated due to interspecies differences. The errors in Table 4 may also be similarly overestimated.

# Discussion

TK information is needed to understand chemical risk posed to human health. This information includes half-life (thalf) and toxicological dose metrics like peak concentration (Cmax) and time-integrated plasma concentration (AUC). Key aspects of TK are chemical-specific and chemical-specific data is unavailable for thousands of chemicals in commerce and the environment. In particular, these TK data gaps exist for chemicals identified in human biological matrices by both targeted (U.S. Centers for Disease Control and Prevention 2012) and non-targeted analyses (Wang et al. 2018). Multiple governments have recognized (Health Canada 2021; National Academies of Sciences 2017; Paini et al. 2020; U.S. Environmental Protection Agency 2021) that a high throughput (chemical-agnostic) TK model parameterized with chemical-specific *in vitro* data (that is, HTTK) can be a powerful tool. Translation of *in vitro* concentrations to putative *in vivo* doses by HTTK facilitates 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 (Sayre et al. 2020; Wambaugh et al. 2018) and *in vitro* HTTK data (Paini et al. 2020; Rotroff et al. 2010; Tonnelier et al. ; Wambaugh et al. 2019; Wetmore et al. 2015; Wetmore et al. 2012). There is an ongoing proliferation of high throughput PBTK models developed to make use of these HTTK *in vitro* parameters to allow HT-PBTK models to make chemical-specific predictions (Armitage et al. 2021; Bernstein et al. 2021; Breen et al. 2021; Geci et al. 2024; Jamei et al. 2009; Kapraun et al. 2022; Linakis et al. 2020). However, several thousand chemicals remain in need of TK info; researchers and regulators may have to rely on machine learning for toxicokinetics to address this gap (Chou and Lin 2023; Di Lascio et al. 2023).

We have characterized the accuracy of HTTK approaches for new chemicals based on structure-based *in silico* predictions. This was accomplished through a collaborative trial of four QSPRs for *in vitro* TK parameters and two additional predictors of *in vivo* TK half-life. This collaborative trial used a database of *in vivo* measured TK 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. We find that for novel compounds a consensus QSPR approach may yield a reasonable result.

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 chemical specific HTTK parameters Clint and fup that are expected by HT-PBTK models. These *in vitro* parameters, of course, are not true TK. Even if QSPR perfectly reproduces an *in vitro* measurement, it is only as good as the *in vitro* assays. However, HTTK assays provide rapid methods for partially characterizing TK (Breen et al. 2021; Wang 2010). Most QSPRs predict plasma binding well (RMSLE 0.03 – 0.07) but have difficulty predicting metabolic clearance (RMSLE 0.37 – 1.28). It is known that some of the QSPRs included the evaluation data in their training sets. Thus, this evaluation is not a true external evaluation, but rather an evaluation of opportunity. Still, it is important to contrast the predictive ability of QSPRs with various training sets (ADMET, for example, has an extensive pharmaceutical library) and characterize how well the QSPRs reproduce the data.

Predictions for the full concentration time-course were then evaluated. Our “level 2” evaluation used the QSPR predicted parameter values within an HT-PBTK model to predict the full TK chemical concentration time course. TK in the absorption and distribution phases are relatively insensitive to the *in vitro* HTTK parameters Clint and fup. For example, distribution depends more on physico-chemical properties and only somewhat on fup for accurate prediction of equilibrium tissue partition coefficients (and, in turn, volume of distribution) (Pearce et al. 2017a). However, the elimination phase of a TK time course is dominated by metabolism and excretion, which are both characterized to some extent by HTTK *in vitro* parameters.

To frame the level 2 evaluation, statistics were also calculated for empirical TK models using parameter optimized using the *in vivo* data. The *in vivo* optimized models were intended to approximate a best case given that the data themselves are noisy. As a worst-case scenario HT-PBTK models were parameterized using randomly selected (incorrect) *in vitro* HTTK parameters, while keeping the physico-chemical parameters correct. The y-randomization approximates helps account for potential correlations within the *in vitro* measured chemical data. During early time points (absorption and distribution) the PBTK models parameterized with QSPRs perform closer to y-randomized predictions than to empirical fits to the data. However, at later time points (elimination phase) the separation between QSPRs and y-randomized predictions grew. The relatively small difference between the average y-randomized error and the QSPR and *in vitro*-based HTTK indicates the great importance of physico-chemical properties to predicting TK.

Here we have found the HT-PBTK model performed similarly between TK QSPRs and *in vitro* measured data. Fits to the *in vivo* data had median RMSLE of 0.68 across all chemicals, while predictions based on *in vitro* HTTK had RMSLE of 1.2 without using any *in vivo* data. RMSLE for individual QSPRs ranged from 1.1 to 1.3. Both *in vitro* data and QSPRs only performed slightly better than using mean prediction when using ten random draws of *in vitro* values for other chemicals (RMSLE 1.3).

An additional consensus prediction was constructed from the various QSPRs. For plasma binding the predictions of the various QSPRs were weighted equally. However, the consensus model for Clint we used the maximum Clint value predicted by any QSPR. Using the maximum Clint corresponds to a more binary description of metabolism – if any of the QSPRs predicts that a compound is rapidly cleared, then we treat it as rapidly cleared in the consensus model. Only if all the QSPRs agree that metabolism is slow do we treat a compound as being slowly metabolized. A consensus prediction using the maximum clearance predicted across all QSPRs had RMSLE 1.1 for the full time-course.

Although TK data for 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 (Kitchin and McArdle 2016). While more than one thousand chemicals have available *in vitro* HTTK measurements for QSPR modeling, 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 94% are within two-fold of the median. Four out of 64 chemicals with i*n vitro* HTTK Clint measured data have no observed clearance (that is, they are metabolically stable) compared with 254 out of 1023 measured Clint values (25 percent) in the httk data set as whole. 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.

QSPR models for metabolism are limited by many factors, including the limitations of the *in vitro* intrinsic clearance model. Since the majority of the *in vitro* Clint assays is based upon hepatocytes suspended in media and losing viability over a four-hour measurement (Rotroff et al. 2010; Shibata et al. 2002), it in unsurprising that the *in vitro* assays underestimate clearance. We currently cannot evaluate the impacts of three-dimensional aspects of chemical structures (including chirality) – all the QSPR models are based on two-dimensional structure descriptors because the data do not exist to train the models otherwise (for example chiral pairs). Further, the data used to train the models are based on human biological material, but the data used for evaluation here are largely drawn from rat, again due to the much wider availability of human than rat *in vitro* HTTK data (Honda et al. 2019; Wetmore et al. 2013).

Geci et al. (2024) recently evaluated HT-PBTK using a large data set of chemical concentration vs. time course data (including the CvTdb) for which AUC and Cmax were estimated via non-compartmental analysis. They found that the error for HT-PBTK parametrized with *in vitro* data was two-fold for Cmax and 1.8-fold for AUC. With *in silico* predictors the best that could be achieved was 2.2-fold for Cmax and 2.4-fold for AUC. Here we observe that empirical one- or two-compartment fits to the CvT data can predicted Cmax with RMSLE 0.7 (5-fold) and AUC with RMSLE 0.62 (4-fold) (Table 5, (Mercado et al. 2024)). A key factor considered by Geci et al. (2024) that was not considered here was intestinal permeability, however given that the majority of the data analyzed here were for rat and there is little reason to expected concordance between rat and human oral absorption (Musther et al. 2014) the source of the discrepancy is unclear. Geci et al. (2024) reports the median, rather than mean error for their results, and for non-normally distributed results (such as log-normal, with a long “tail” of larger values) the mean can be much larger than the median. *In vivo* data are often-log-normally distributed (CAN and LOG 2001). We have focused on the RMSLE here because it provides an estimate of the prediction error for novel chemicals. In model evaluation literature there is significant discussion of the appropriate model evaluation statistic, with mean absolute error (MAE) often preferred for scaling with the magnitude of the prediction (errors are weighted equally) while root mean squared error (RMSE) being preferred for indicating the standard deviation for normally distributed predictions (Chai and Draxler 2014; Willmott and Matsuura 2005). RMSLE provides the advantageous of both, by scaling with the magnitude of prediction (because the data are logarithmic) and allowing statistical interpretation as the coefficient of variation.

Geci et al. (2024) is consistent with (Wang 2010) which found that the SimCYP model, parameterized with *in vitro* data, could predict AUC within 2.3-fold when evaluated across 54 pharmaceuticals. Wambaugh et al. (2018) found evaluated HTTK-based IVIVE for TK using just over forty chemicals including pharmaceuticals and non-pharmaceuticals. Wambaugh et al. (2018) found Cmax could be predicted with RMSLE 2.2 and that AUC could be predicted with RMSLE 2.0. As summarized in Table 5, here we found that Cmax could be predicted with RMSLE 0.82 with *in vitro* data and RMSLE 0.81 with the consensus QSPR. AUC could be predicted with RMSLE 1.3 using *in vitro* data and 1.1 using the consensus QSPR. We speculate that differences relatively to the literature are due to 1) curation of the CvT data by requiring that they can be fit to empirical models (Geci et al. 2024), 2) the inclusion of many non-pharmaceuticals (Wang 2010), and 3) the inclusion of more chemicals (Wambaugh et al. 2018). By design the ADME properties of pharmaceuticals are often constrained to have shorter half-lives than occur for non-pharmaceuticals – this means that the distribution of measured time points may be more sparse for non-pharmaceuticals (Mercado et al. 2024).

Since we are using rat *in vivo* data to characterize the predictive ability of human QSPRs for HTTK, we attempted to identify the error for interspecies differences in HTTK. Using chemicals with Clint and fup measured both in human and rat, we found that predictions of Cmax had median RMSLE of 0.239 across chemicals when human data were used instead of rat. AUC had RMSLE of 0.307. If these errors were the result of uncorrelated factors, which they are not, then one could subtract the interspecies HTTK error estimate to get a better estimate of how well the human QSPRs predict in vivo data. The observed error for AUC was 1.3 (in Table 5), while the interspecies error was estimated to be 0.307. This means that the values estimated here may more consistent with Wang (2010) when interspecies differences are taken into account. The estimated errors for the time course data (Table X) might similarly be overestimated by 1.5 to two times (0.2 to 0.3 RMSLE).

Among the summary TK statistics (level 3) evaluated, those involved in the elimination phase were hardest to predict. The related quantities AUC, CLtot, and half-life are more challenging than Cmax and Vd. 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 0.46 while using *in vitro* values for random (incorrect) chemicals had an AUC RMSLE of 1.9. Using chemical-specific *in vitro* data predicted AUC with RMSLE 1.3, while QSPRs ranged in accuracy from 1.3-1.5. The consensus prediction using the maximum clearance predicted across all QSPRs predicted AUC with RMSLE 1.1 – this is better than using the *in vitro* measured data for the evaluation chemicals. Using the consensus model, the total clearance CLtot RMSLE is 1.6. For Cmax *in vivo* fits gave RMSLE 0.56 and random chemicals gave 1. Chemical-specific *in vitro* data had RMSLE 0.85 and QSPRs ranged from 0.79-0.99, with the consensus being 0.87. Overall predictions of Vd had RMSLE ranging between 1.2-1.4, with the random model yielding 1.1. Here we have assumed that Fbio only depends on first-pass metabolism, as informed by Clint. This bioavailability assumption does not appear to explain the observed discrepancies with *in vivo* data. The uncertainty analysis (Figure 8) confirmed that improving HTTK estimates of the elimination phase would be most likely to enhance prediction accuracy.

QSPRs trained to true human half-life values (such as QSARINS and IFS-QSAR) are trained to more chemicals (the totality of chemicals with known half-lives rather than just those with in vitro HTTK data) and capture whole body clearance (rather than just the hepatic metabolism, glomerular filtration, and exhalation of HTTK). We found that using the half-life predictions with the Vd predicted by the consensus model gave more accurate predictions of whole-body clearance (and therefore of Css). As half-lives, much of the functionality of HTTK not discussed here cannot be used. These aspects of HTTK include the ability to simulate human variability to identify sensitive populations (Breen et al. 2022) and the ability to propagate chemical-specific measurement uncertainty (Wambaugh et al. 2019). However, since the half-life models proved superior estimates, it may be worth exploring converting them to corresponding in vitro HTTK parameters and integrating them more fully with other HTTK methods.

For a novel chemical it may be safer to assume that the prediction errors will be closer to the conservative values estimated here. This is supported by the observed accuracy for the comparing the predicted full concentration vs. time curves against the observations, which are similar accurate to the predictions of summary TK statistics Cmax and AUC that have been analyzed elsewhere. We do not necessarily think that the improved ability of consensus modeling is sole influence of any one model, rather that consensus QSPRs represent the “wisdom of the crowd” (Mansouri et al. 2016; Pradeep et al. 2016). By using the maximum predicted clearance, if any one QSPR predicts clearance then chemical may be more likely to be metabolized. We hope 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

Akaike, H. Information theory and an extension of the maximum likelihood principle. Selected papers of hirotugu akaike: Springer; 1998

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

Arnot, J.A.; Brown, T.N.; Wania, F. Estimating screening-level organic chemical half-lives in humans. Environmental science & technology 2014;48:723-730

Baumer, B.; Udwin, D. R markdown. Wiley Interdisciplinary Reviews: Computational Statistics 2015;7:167-177

Bell, S.M.; Chang, X.; Wambaugh, J.F.; Allen, D.G.; Bartels, M.; Brouwer, K.L.R.; Casey, W.M.; Choksi, N.; Ferguson, S.S.; Fraczkiewicz, G.; Jarabek, A.M.; Ke, A.; Lumen, A.; Lynn, S.G.; Paini, A.; Price, P.S.; Ring, C.; Simon, T.W.; Sipes, N.S.; Sprankle, C.S.; Strickland, J.; Troutman, J.; Wetmore, B.A.; Kleinstreuer, N.C. In vitro to in vivo extrapolation for high throughput prioritization and decision making. Toxicology in Vitro 2018;47:213-227

Bernstein, A.S.; Kapraun, D.F.; Schlosser, P.M. 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 2021;

Breen, M.; Ring, C.L.; Kreutz, A.; Goldsmith, M.-R.; Wambaugh, J.F. High-throughput PBTK models for in vitro to in vivo extrapolation. Expert Opinion on Drug Metabolism & Toxicology 2021;17:903-921

Breen, M.; Wambaugh, J.F.; Bernstein, A.; Sfeir, M.; Ring, C.L. Simulating Toxicokinetic Variability to Identify Susceptible and Highly Exposed Populations. Journal of Exposure Science & Environmental Epidemiology 2022;

CAN, W.; LOG, O. Log-normal Distributions across the Sciences: Keys and Clues. BioScience 2001;51

Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development 2014;7:1247-1250

Chirico, N.; Bertato, L.; Papa, E. In Vitro Biotransformation Prediction-Suite (IVBP-Suite). 2021;

Chou, W.-C.; Lin, Z. Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. Toxicological Sciences 2023;191:1-14

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

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

Dawson, D.E.; Ingle, B.L.; Phillips, K.A.; Nichols, J.W.; Wambaugh, J.F.; Tornero-Velez, R. Designing QSARs for Parameters of High-Throughput Toxicokinetic Models Using Open-Source Descriptors. Environmental science & technology 2021;55:6505-6517

Di Lascio, E.; Gerebtzoff, G.; Rodríguez-Pérez, R. Systematic evaluation of local and global machine learning models for the prediction of ADME properties. Molecular pharmaceutics 2023;20:1758-1767

Fenton, S.E.; Ducatman, A.; Boobis, A.; DeWitt, J.C.; Lau, C.; Ng, C.; Smith, J.S.; Roberts, S.M. Per- and Polyfluoroalkyl Substance Toxicity and Human Health Review: Current State of Knowledge and Strategies for Informing Future Research. Environmental Toxicology and Chemistry 2021;40:606-630

Geci, R.; Gadaleta, D.; de Lomana, M.G.; Ortega-Vallbona, R.; Colombo, E.; Serrano-Candelas, E.; Paini, A.; Kuepfer, L.; Schaller, S. Systematic evaluation of high-throughput PBK modelling strategies for the prediction of intravenous and oral pharmacokinetics in humans. Archives of Toxicology 2024;

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

Honda, G.S.; Pearce, R.G.; Pham, L.L.; Setzer, R.W.; Wetmore, B.A.; Sipes, N.S.; Gilbert, J.; Franz, B.; Thomas, R.S.; Wambaugh, J.F. Using the concordance of in vitro and in vivo data to evaluate extrapolation assumptions. PLOS ONE 2019;14:e0217564

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

Kapraun, D.F.; Sfeir, M.; Pearce, R.; Davidson, S.; Lumen, A.; Dallmann, A.; Judson, R.; Wambaugh, J.F. Evaluation of a Rapid, Generic Human Gestational Dose Model. 2022

Kavlock, R.J.; Bahadori, T.; Barton-Maclaren, T.S.; Gwinn, M.R.; Rasenberg, M.; Thomas, R.S. Accelerating the Pace of Chemical Risk Assessment. Chemical Research in Toxicology 2018;31:287-290

Kirman, C.R.; Aylward, L.L.; Wetmore, B.A.; Thomas, R.S.; Sochaski, M.; Ferguson, S.S.; Csiszar, S.A.; Jolliet, O. Quantitative property–property relationship for screening-level prediction of intrinsic clearance: a tool for exposure modeling for high-throughput toxicity screening data. Applied In Vitro Toxicology 2015;1:140-146

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

Linakis, M.W.; Sayre, R.R.; Pearce, R.G.; Sfeir, M.A.; Sipes, N.S.; Pangburn, H.A.; Gearhart, J.M.; Wambaugh, J.F. Development and Evaluation of a High Throughput Inhalation Model for Organic Chemicals Journal of Exposure Science & Environmental Epidemiology 2020;30:866-877

Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A.; Grulke, C.M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenat, E.; Muratov, E.; Wedebye, E.B.; Grisoni, F.; Mangiatordi, G.F.; Incisivo, G.M.; Hong, H.; Ng, H.W.; Tetko, I.V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; N. G. Nikolov; Nicolotti; Andersson, P.L.; Zang, Q.; Politi, R.; Beger, R.D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S.A.; Slavov, S.; Hu, X.; Judson, R.S. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environmental Health Perspectives 2016;124:1023-1033

Mansouri, K.; Chang, X.; Allen, D.; Judson, R.; Williams, A.; Kleinstreuer, N. OPERA models for ADME properties and toxicity endpoint. Society of Toxicology Annual Meeting; 2021

Mansouri, K.; Grulke, C.M.; Judson, R.S.; Williams, A.J. OPERA models for predicting physicochemical properties and environmental fate endpoints. Journal of Cheminformatics 2018;10:10

Mercado, G.P.; Cook, C.; Albrecht, L.; Cary, G.; Edwards, B.; Haggard, D.; Hanley, N.; Hughes, M.F.; Jarnagin, A.; Kodavanti, T.D.; Koral-Bexell, E.; Kreutz, A.; Ngo, M.; Patullo, C.; Sayre, R.; Sharma, B.; Wall, J.; Yamazaki, H.; Wambaugh, J.; Ring, C. An informatics framework for toxicokinetics using the CvT database and invivoPKfit. Society of Toxicology Annual Meeting. Salt Lake City, Utah; 2024

Musther, H.; Olivares-Morales, A.; Hatley, O.J.; Liu, B.; Hodjegan, A.R. Animal versus human oral drug bioavailability: do they correlate? European Journal of Pharmaceutical Sciences 2014;57:280-291

National Academies of Sciences, E., and Medicine,. Using 21st Century Science to Improve Risk-Related Evaluations ed^eds. Washington, DC: The National Academies Press; 2017

National Research Council. Risk Assessment in the Federal Government: Managing the Process. Washington (DC): National Academies Press; 1983

O'Flaherty, E.J. Toxicants and drugs: kinetics and dynamics ed^eds: John Wiley & Sons; 1981

Paini, A.; Cole, T.; Meinero, M.; Carpi, D.; Deceuninck, P.; Macko, P.; Palosaari, T.; Sund, J.; Worth, A.; Whelan, M. EURL ECVAM in vitro hepatocyte clearance and blood plasma protein binding dataset for 77 chemicals. in: European Commission J.R.C.J., ed; 2020

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

Patlewicz, G.; Richard, A.M.; Williams, A.J.; Grulke, C.M.; Sams, R.; Lambert, J.; Noyes, P.D.; DeVito, M.J.; Hines, R.N.; Strynar, M. A chemical category-based prioritization approach for selecting 75 per-and polyfluoroalkyl substances (PFAS) for tiered toxicity and toxicokinetic testing. Environmental health perspectives 2019;127:014501

Pearce, R.G.; Setzer, R.W.; Davis, J.L.; Wambaugh, J.F. Evaluation and calibration of high-throughput predictions of chemical distribution to tissues. Journal of Pharmacokinetics and Pharmacodynamics 2017a;44:549-565

Pearce, R.G.; Setzer, R.W.; Strope, C.L.; Sipes, N.S.; Wambaugh, J.F. httk: R package for high-throughput toxicokinetics. Journal of Statistical Software 2017b;79:1-26

Peyret, T.; Poulin, P.; Krishnan, K. A unified algorithm for predicting partition coefficients for PBPK modeling of drugs and environmental chemicals. *Toxicology and Applied Pharmacology* 2010;249:197-207

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

Pradeep, P.; Povinelli, R.J.; White, S.; Merrill, S.J. An ensemble model of QSAR tools for regulatory risk assessment. Journal of cheminformatics 2016;8:1-9

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

Richard, A.M.; Judson, R.S.; Houck, K.A.; Grulke, C.M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M.T.; Wambaugh, J.F. ToxCast chemical landscape: Paving the road to 21st century toxicology. Chemical Research in Toxicology 2016;29:1225-1251

Rotroff, D.M.; Wetmore, B.A.; Dix, D.J.; Ferguson, S.S.; Clewell, H.J.; Houck, K.A.; LeCluyse, E.L.; Andersen, M.E.; Judson, R.S.; Smith, C.M.; Sochaski, M.A.; Kavlock, R.J.; Boellmann, F.; Martin, M.T.; Reif, D.M.; Wambaugh, J.F.; Thomas, R.S. Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. Toxicological Sciences 2010;117:348-358

Rydberg, P.; Gloriam, D.E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. ACS Medicinal Chemistry Letters 2010;1:96-100

Sayre, R.R.; Wambaugh, J.F.; Grulke, C.M. Database of pharmacokinetic time-series data and parameters for 144 environmental chemicals. Scientific Data 2020;

Schmidt, C.W. TSCA 2.0: A new era in chemical risk management. National Institute of Environmental Health Sciences; 2016

Schmitt, W. General approach for the calculation of tissue to plasma partition coefficients. Toxicology in Vitro 2008;22:457-467

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

Sipes, N.S.; Wambaugh, J.F.; Pearce, R.; Auerbach, S.S.; Wetmore, B.A.; Hsieh, J.-H.; Shapiro, A.J.; Svoboda, D.; DeVito, M.J.; Ferguson, S.S. An Intuitive Approach for Predicting Potential Human Health Risk with the Tox21 10k Library. Environmental Science & Technology 2017;51:10786-10796

Sobus, J.R.; Tan, Y.-M.; Pleil, J.D.; Sheldon, L.S. A biomonitoring framework to support exposure and risk assessments. Science of the Total Environment 2011;409:4875-4884

Tice, R.R.; Austin, C.P.; Kavlock, R.J.; Bucher, J.R. Improving the human hazard characterization of chemicals: a Tox21 update. Environmental Health Perspectives 2013;121:756

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

U.S. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey. 2012

U.S. Environmental Protection Agency. U.S. EPA. A Proof-of-Concept Case Study Integrating Publicly Available Information to Screen Candidates for Chemical Prioritization under TSCA. Washington, DC; 2021

Wambaugh, J.F.; Hughes, M.F.; Ring, C.L.; MacMillan, D.K.; Ford, J.; Fennell, T.R.; Black, S.R.; Snyder, R.W.; Sipes, N.S.; Wetmore, B.A.; Westerhout, J.; Setzer, R.W.; Pearce, R.G.; Simmons, J.E.; Thomas, R.S. Evaluating in vitro-in vivo extrapolation of toxicokinetics. Toxicological Sciences 2018;163:152-169

Wambaugh, J.F.; Rager, J.E. Exposure forecasting–ExpoCast–for data-poor chemicals in commerce and the environment. Journal of exposure science & environmental epidemiology 2022;32:783-793

Wambaugh, J.F.; Wetmore, B.A.; Ring, C.L.; Nicolas, C.I.; Pearce, R.; Honda, G.; Dinallo, R.; Angus, D.; Gilbert, J.; Sierra, T.; Badrinarayanan, A.; Snodgrass, B.; Brockman, A.; Strock, C.; Setzer, W.; Thomas, R.S. Assessing Toxicokinetic Uncertainty and Variability in Risk Prioritization. Toxicological Sciences 2019;172:235-251

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

2018;126:077009

Wang, Y.-H. 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 2010;38:1094-1104

Warnes, G.R.; Bolker, B.; Bonebakker, L.; Gentleman, R.; Huber, W.; Liaw, A.; Lumley, T.; Maechler, M.; Magnusson, A.; Moeller, S.; Schwartz, M.V., Bill. gplots: various R programming tools for plotting data. R package version 3.1.3.1. Computer software; 2024

Waters, N.J.; Jones, R.; Williams, G.; Sohal, B. Validation of a rapid equilibrium dialysis approach for the measurement of plasma protein binding. Journal of Pharmaceutical Sciences 2008;97:4586-4595

Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences 1988;28:31-36

Wetmore, B.A. Quantitative in vitro-to-in vivo extrapolation in a high-throughput environment. Toxicology 2015;332:94-101

Wetmore, B.A.; Wambaugh, J.F.; Allen, B.; Ferguson, S.S.; Sochaski, M.A.; Setzer, R.W.; Houck, K.A.; Strope, C.L.; Cantwell, K.; Judson, R.S.; LeCluyse, E.; Clewell, H.J.; Thomas, R.S.; Andersen, M.E. Incorporating High-Throughput Exposure Predictions With Dosimetry-Adjusted In Vitro Bioactivity to Inform Chemical Toxicity Testing. Toxicological Sciences 2015;148:121-136

Wetmore, B.A.; Wambaugh, J.F.; Ferguson, S.S.; Li, L.; Clewell, H.J., 3rd; Judson, R.S.; Freeman, K.; Bao, W.; Sochaski, M.A.; Chu, T.M.; Black, M.B.; Healy, E.; Allen, B.; Andersen, M.E.; Wolfinger, R.D.; Thomas, R.S. Relative impact of incorporating pharmacokinetics on predicting in vivo hazard and mode of action from high-throughput in vitro toxicity assays. Toxicological Sciences 2013;132:327-346

Wetmore, B.A.; Wambaugh, J.F.; Ferguson, S.S.; Sochaski, M.A.; Rotroff, D.M.; Freeman, K.; Clewell, H.J., 3rd; Dix, D.J.; Andersen, M.E.; Houck, K.A.; Allen, B.; Judson, R.S.; Singh, R.; Kavlock, R.J.; Richard, A.M.; Thomas, R.S. Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. Toxicological Sciences 2012;125:157-174

Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 2005;30:79-82

# Figures

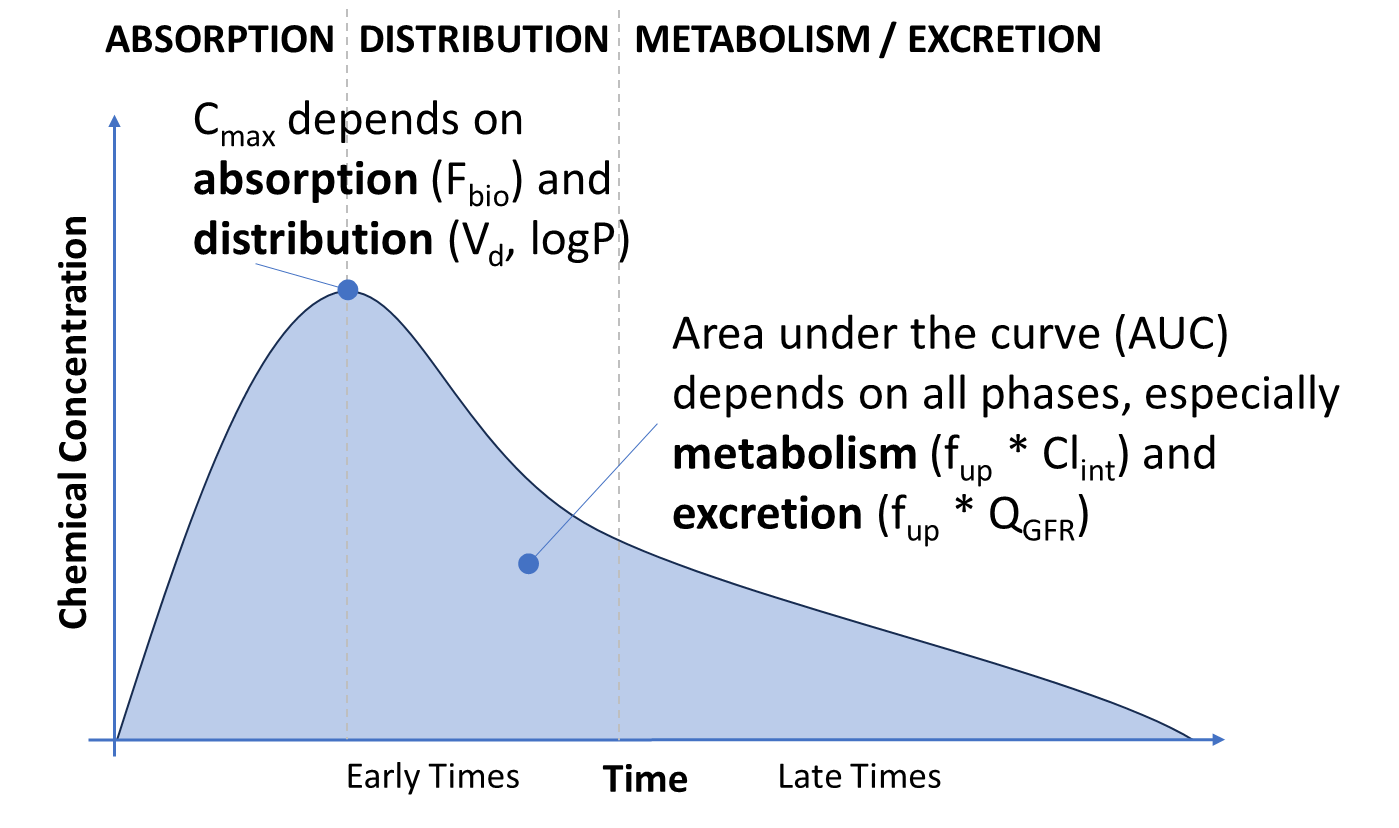


Figure 1: Conceptual illustration of toxicokinetic concentration vs. time (CvT) data.

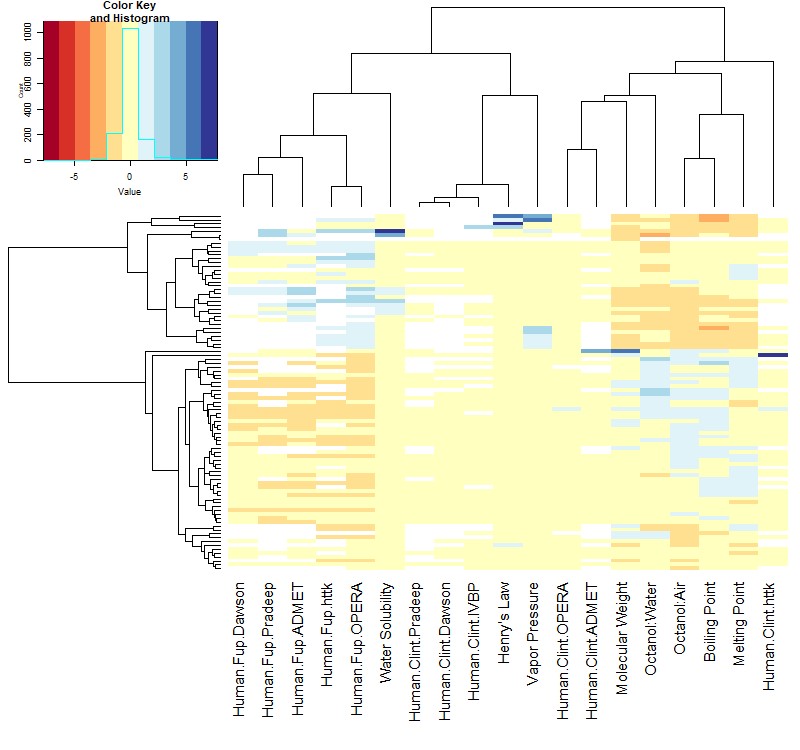


Figure 2: Columns in this heatmap indicate the physico chemical properties and measured/predicted values for in vitro TK (intrinsic hepatic clearance, Clint, and fraction unbound in plasma, fup). Each row corresponds to one of the l92 chemicals. The in vitro TK measurements (“Human.Clint.InVitro” and “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.

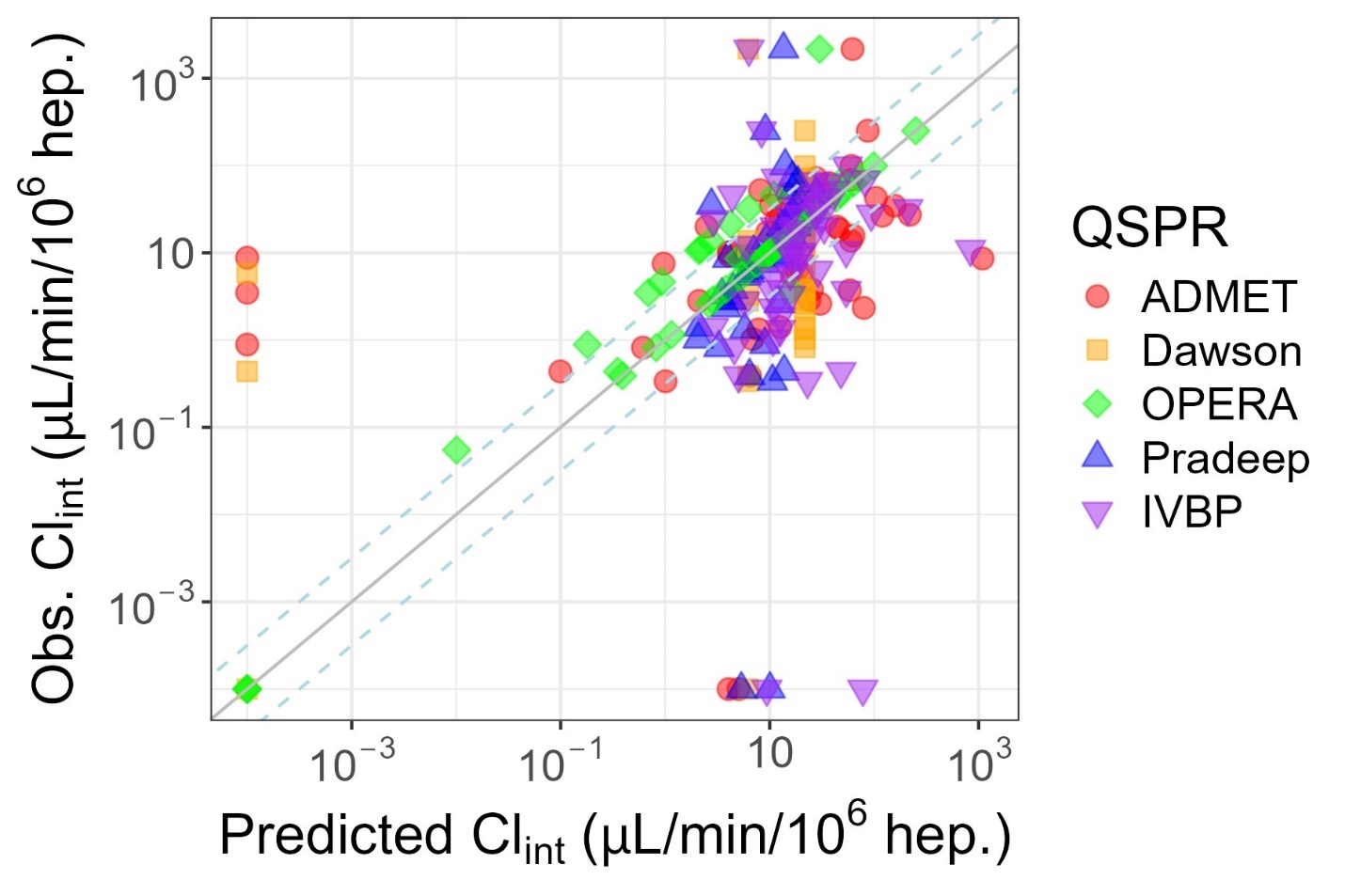


Figure 3: Evaluation of Predictions for Intrinsic Hepatic Clearance (Clint). Zero values were plotted at 10-1, the solid line indicates identity (1:1) while the dashed lines indicate 3.2-fold difference.

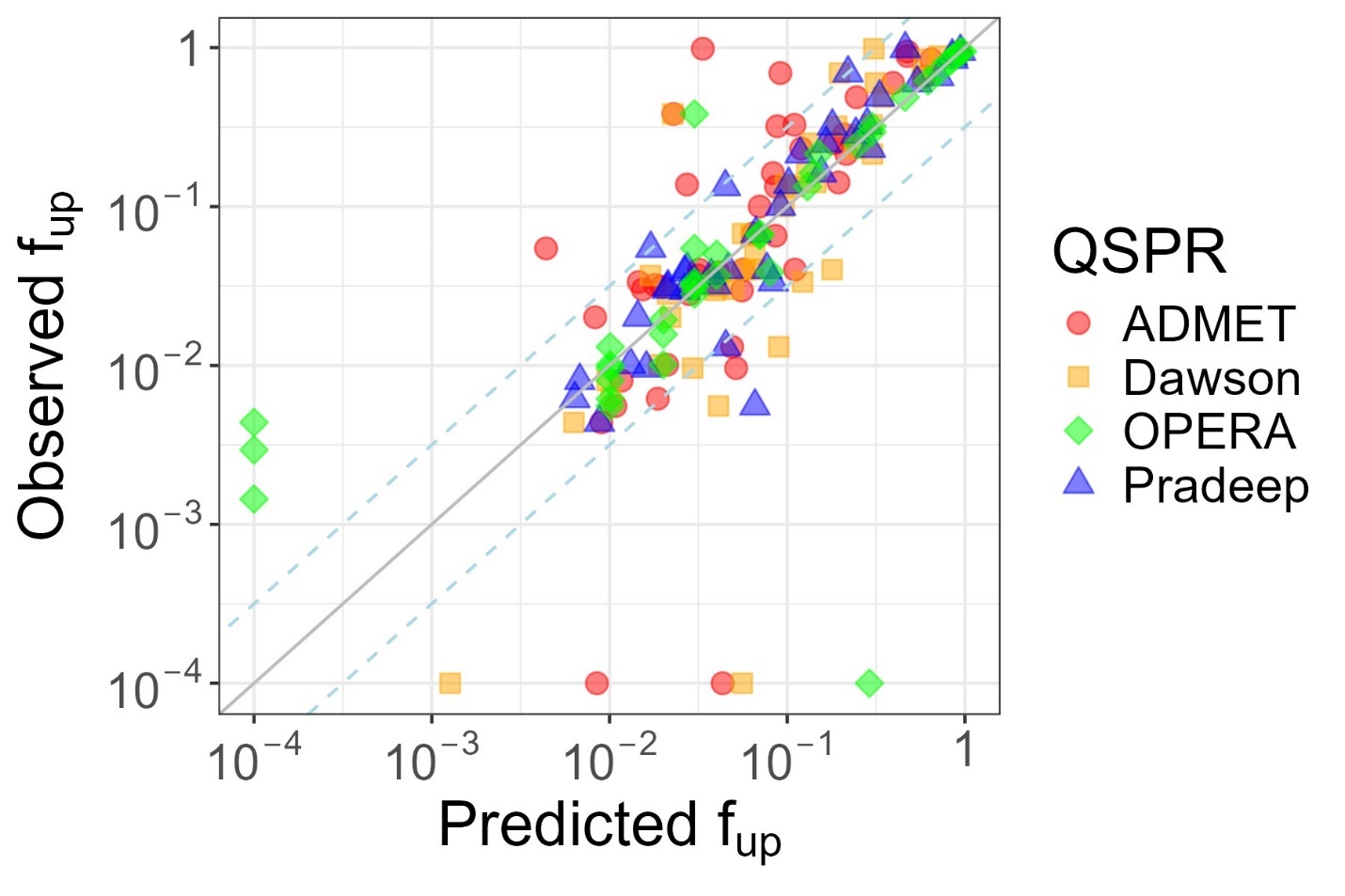


Figure 4: Evaluation of Predictions for Fraction Unbound in Plasma (fup). Zero values were plotted at 10-4, the solid line indicates identity (1:1) while the dashed lines indicate 3.2-fold difference.

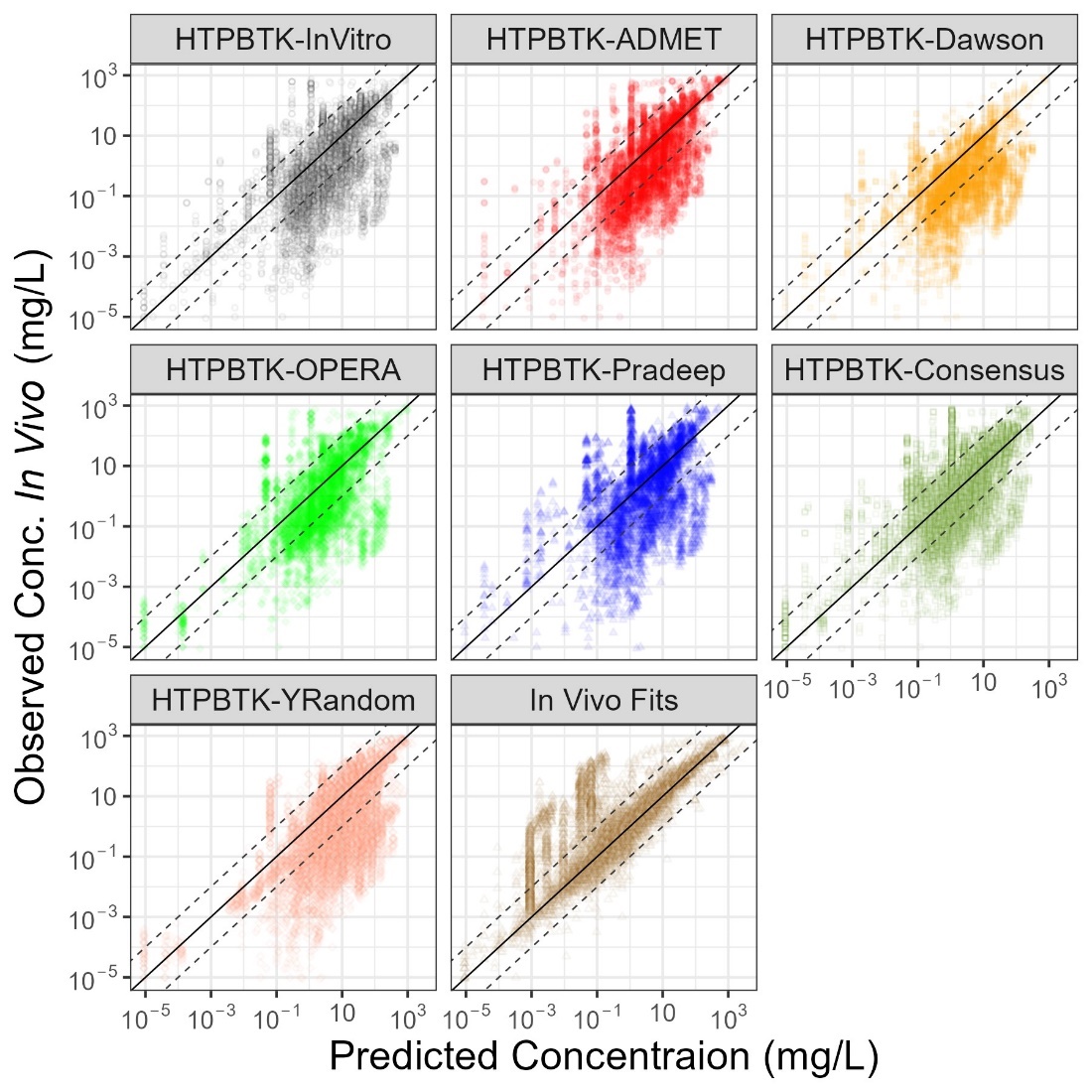


Figure 5: Comparison of in vivo measured chemical concentration vs. time (CvT) data (Sayre et al. 2020) vs. predictions for empirical models fit to the data (“In vivo Fits”), and predictions for a PBTK model (“HTTK”) parameterized with chemical specific values either measured in vitro (“HTTK-InVitro”) 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.

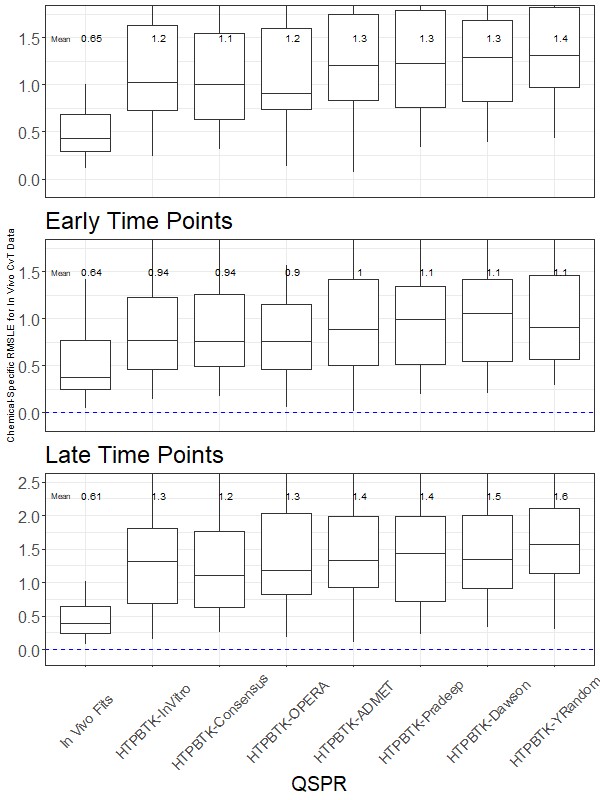


Figure 6: Chemical-Specific RMSLE for In vivo CvT Data. The upper and lower extent of the box for each model indicates the 25th to 75th quantiles, the mid-line indicates the median (50th quantile) and vertical line indicates 1.5x the range of the box.

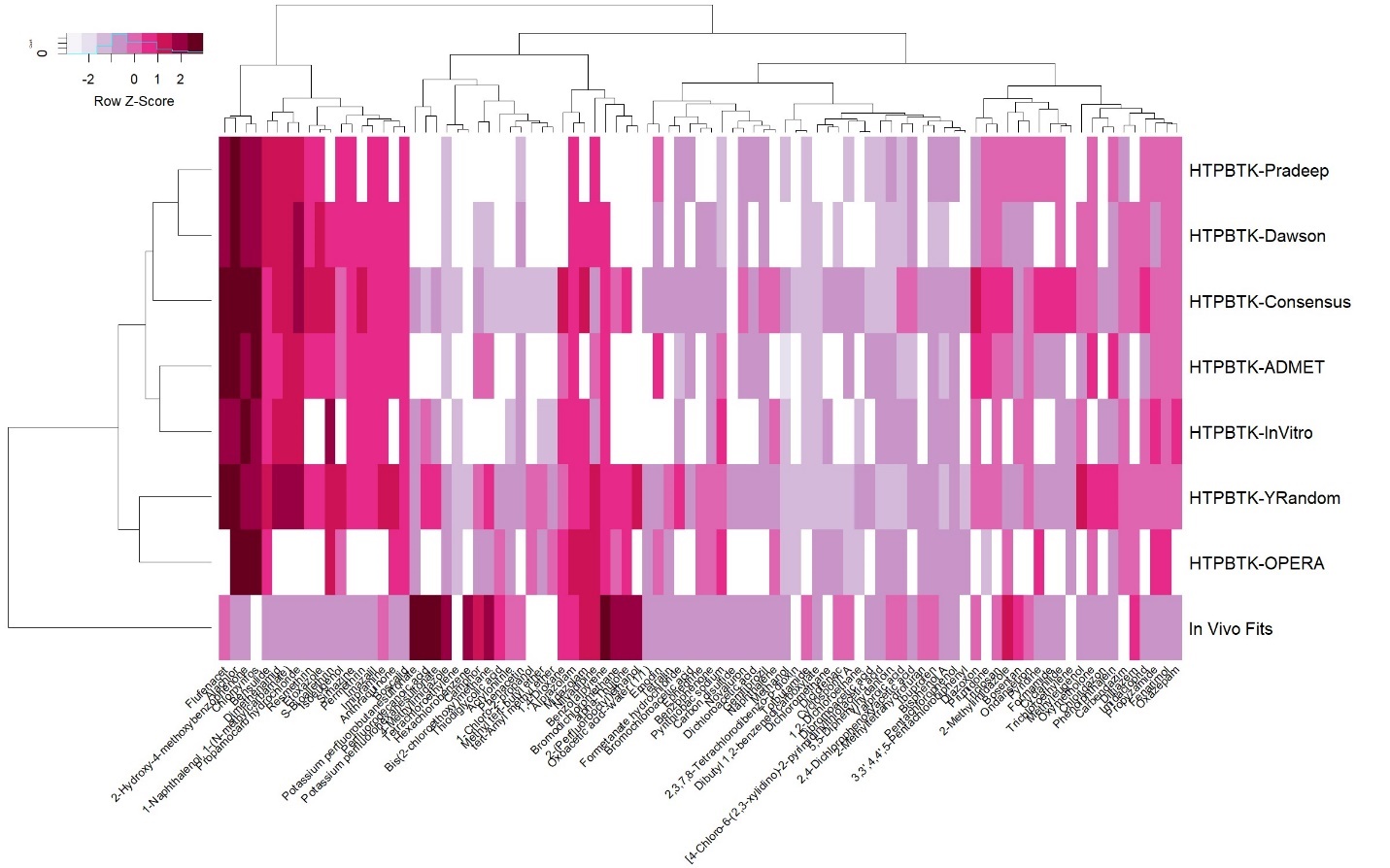


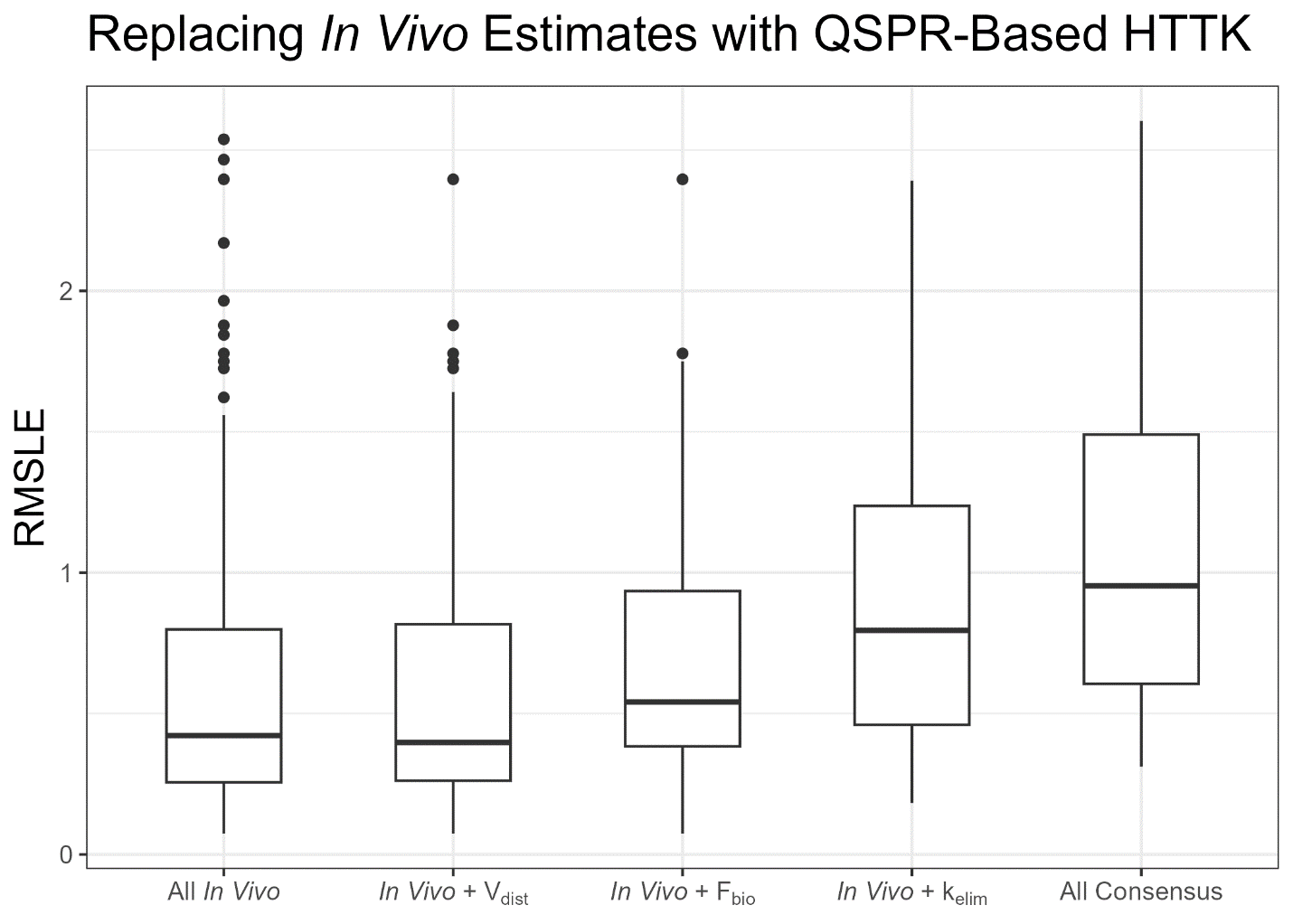
Figure 7: Values in this heatmap are the root mean squared log10 error (RMSLE) by chemical and predictor (that is, lighter indicates better predictive ability). The columns indicate different evaluation chemicals. The rows give the different prediction methods: The empirical fits to the data are given by “In vivo Fits”. All other values are calculated using the HT-PBTK model and either measured values “In vitro”, y-randomized measured values (“Y-Random”) or the various QSPRS. 

Figure 8 Box-and-whisker plot indicating median, interquartile, 95%, and outlier RMSLE on a per chemical basis. The lowest median error is when all parameters for an empirical TK model are derived from the in vivo time course data (“All In Vivo”). Substituting values predicted from the Consensus QSPR with HT-PBTK indicates which aspects of TK are the largest contributor to error. “All Consensus” indicates the error when all parameters are estimated with the Consensus QSPR and HT-PBTK.

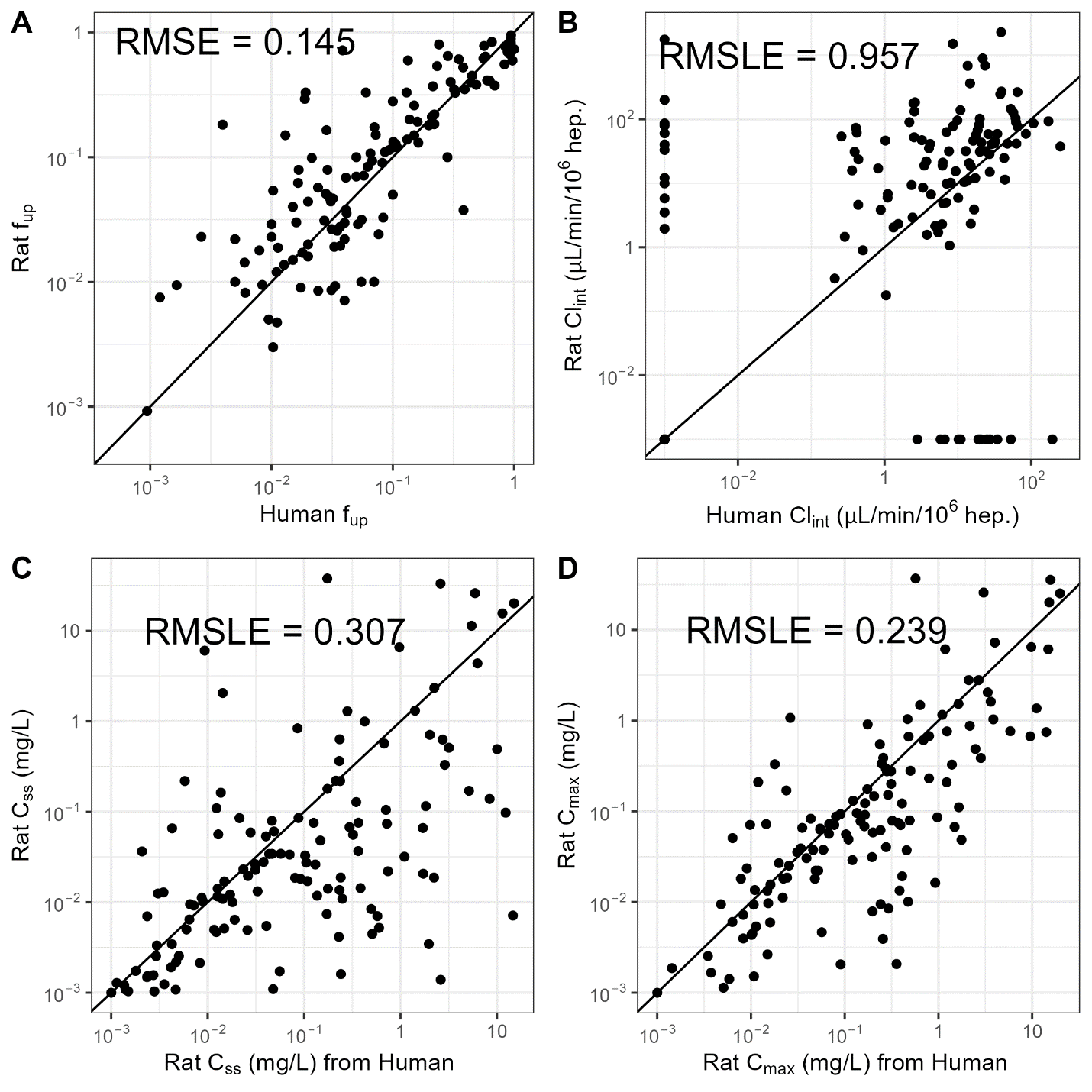
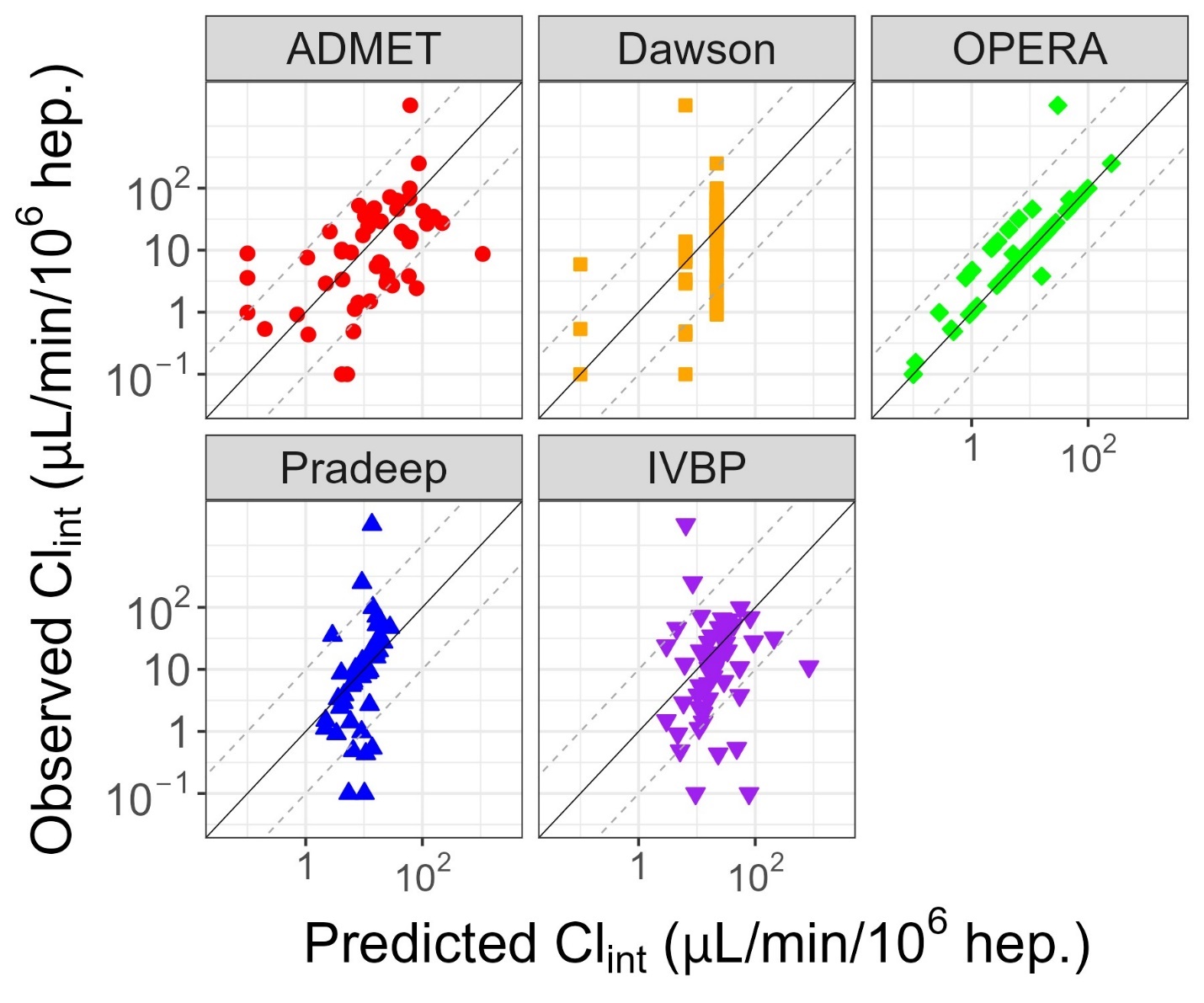


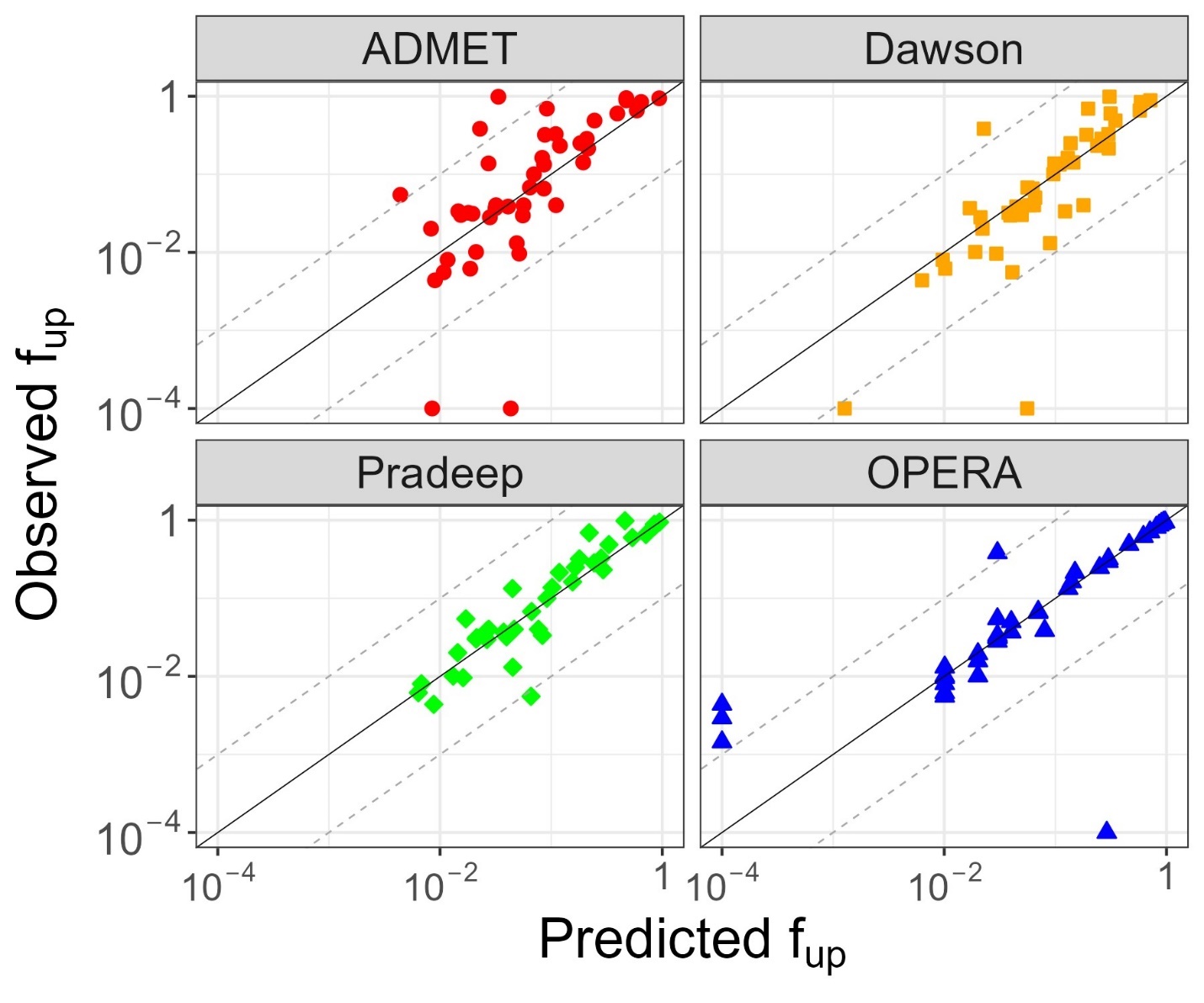
Figure 9 Interspecies concordance of HTTK for chemicals with parameters measured in both rat and human: (panel a) fraction unbound in plasma measured in vitro, (panel b) intrinsic hepatic clearance measured in vitro, (panel c) predicted steady-state plasma concentration predicted with HT-PBTK, and (panel d) predicted peak plasma concentration predicted with HT-PBTK. In panel b, occurrences of zero clearance are plotted at 10-3 so that they appear on the logarithmic scale.

# Supplemental Figures

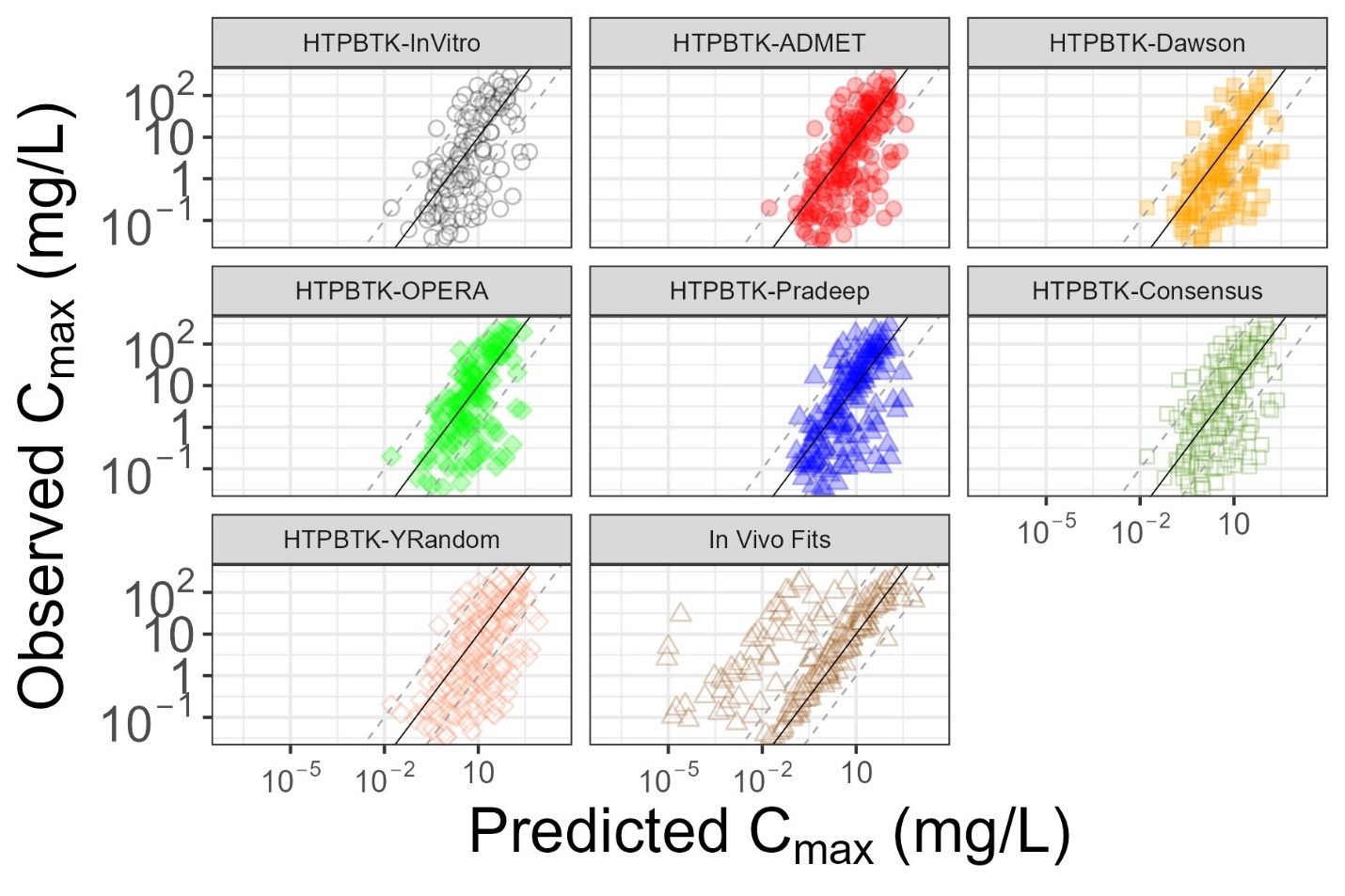
Supplemental Materials are available at: <https://github.com/USEPA/CompTox-ExpoCast-HTTKQSPRs>



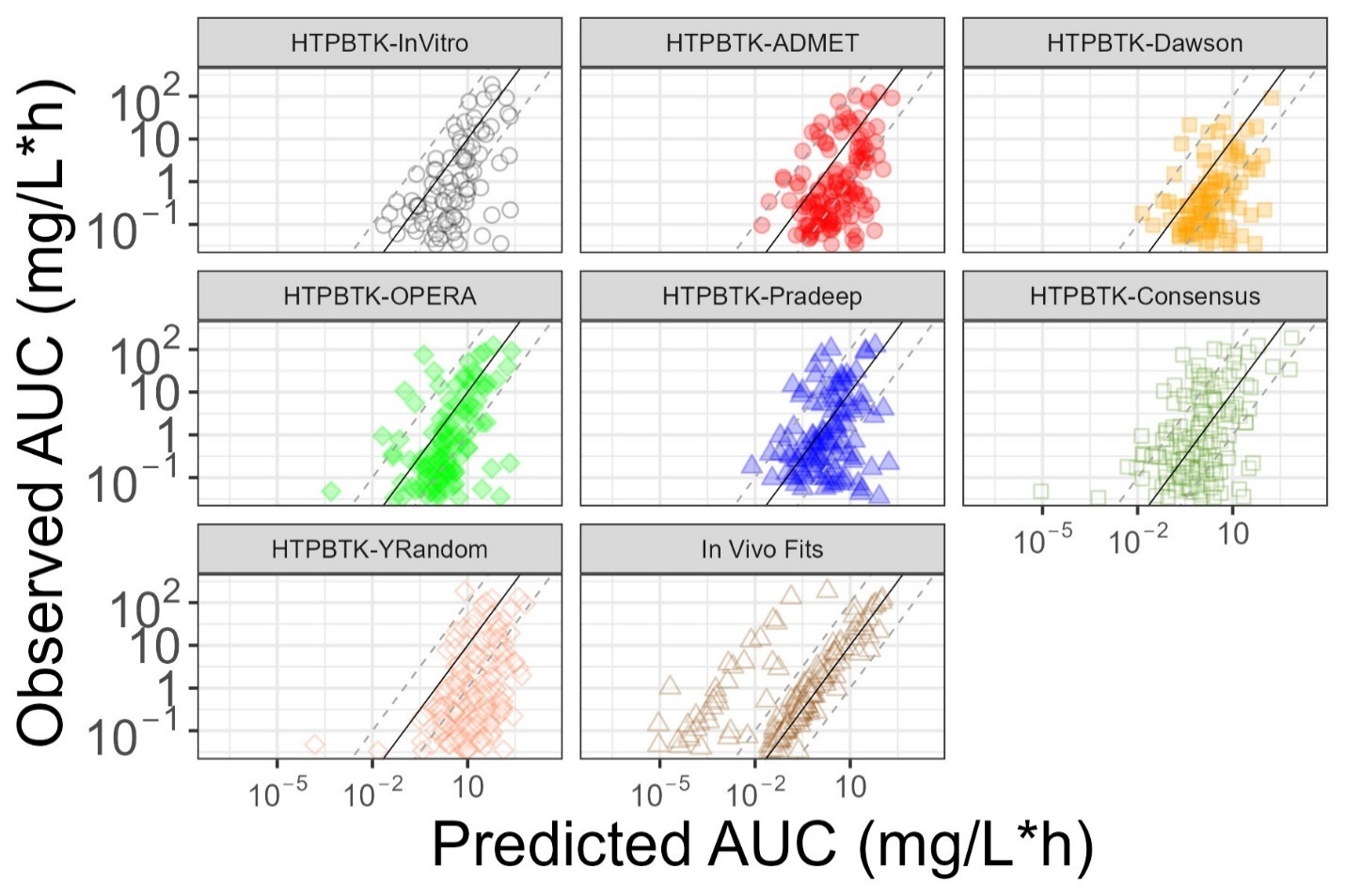
Supplemental Figure 1: Evaluation of Predictions for Intrinsic Hepatic Clearance (Clint). Zero values were plotted at 10-1, the solid line indicates identity (1:1) while the dashed lines indicate 3.2-fold difference.



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 3: 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 4: 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 Materials are available at: <https://github.com/USEPA/CompTox-ExpoCast-HTTKQSPRs>

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 invivoPKfit

Supplemental Table 45: 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 56: Chemical-specific root mean square log10 errors for the full TK concentration time course data by QSPRs*SupTable-RMSLEbyChem.txt*

Supplemental Table6 7: Level 3 Predictions

SupTable-Level3.xlsx