Designing QSARs for parameters of high throughput toxicokinetic models using open-source descriptors

*Daniel Dawson1, Brandall L. Ingle2, Katherine A. Phillips1, John W. Nichols1, John F. Wambaugh1, Rogelio Tornero-Velez1\**

1U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709

2U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709

S1. Supplemental Methods

S1.1 Hepatic Clearance Data Processing

S1.1.1: Selecting Chemicals with Only 1st Order Clearance Dynamics

 Chemicals in the ToxCast screening program, hereafter known as the “ToxCast Dataset” included chemicals from a variety of domains that were reported in the literature.6, 7, 15, 16 Values for *Clint* collected from ToxCast were generated by measuring parent chemical depletion over 2-4 hours (at initial concentrations of 1 and 10 µM) in an assay performed using human hepatocytes. 6, 15-17 Insufficient information was available in these assays to estimate non-linear clearance dynamics (e.g., Michaelis-Menton behavior). Therefore, we attempted to include only chemicals for which data appeared to be collected under first-order conditions (i.e., linear) conditions by considering chemicals for which the clearance rate measured at 1 µM was less than or equal to the rate at 10 µM (895). Then, the average value of these two values was taken.

S1.1.2: Adjusting for Chemical Binding

 To account for the reduction in *Clint*rates due to chemical binding (e.g., protein, lipids) in microsomal and hepatic assays, values from both CHEMBL and ToxCast were adjusted by dividing by a predicted fraction of the chemical unbound in the assays. These values were calculated using methods by Kilford et al.1 based on the lipophilicity characteristics of the chemical using the following equations:

 (1)

for microsomal assays, or

 (2)

for hepatic assays. In equation 1, the term P equals the protein content of the assay, while logD74 is the log of the octanal/water coefficient at pH 7.4. In equation 2, the term VR is the ratio of cells to incubation volume for assay. Values of P and VR for each assay are likely variable but uncertain, and thus were assumed to be constants equal to 1 mg/ml and 0.005, respectively following Kilford et al.1

S1.1.3. *Clint* training and test set assembly

 To construct and assess models for *Clint*, the overall dataset was split into a training set consisting of both ToxCast and ChEMBL data and two independent validation sets consisting of chemicals from ToxCast or ChEMBL. There were two primary considerations in assembling the training set. First, equal weight was given to the inclusion of the ToxCast and ChEMBL chemical datasets while retaining a reasonable test set for the ToxCast chemicals, the smaller of the two datasets. Because non-pharmaceutical chemicals are under-represented in previous QSAR models of *Clint*, it was decided that utilizing a smaller training set with greater diversity in chemical space was preferable to using a training set dominated by pharmaceutical data. Second, equal representation was given to each bin (very slow, slow, fast, very fast). These two priorities resulted in the random selection of a training set consisting of 1600 chemicals, including 804 ChEMBL chemicals and 796 ToxCast Chemicals. While bins were generally determined based on approximate quartiles of *in vitro* data, the exact boundaries for each classification bin were adjusted to provide a uniform distribution of ~400 chemicals from the total training dataset in each category (Fig S1.1). Remaining chemicals in each dataset were included in test sets, including a set of 99 ToxCast chemicals (representing about 11% of the ToxCast dataset with available *Clint* data), and a set of 2112 ChEMBL chemicals (representing about 72% of the ChEMBL dataset with available *Clint* data). While test sets were not explicitly balanced by classification bin, both sets consisted of least 8% of each bin, with this lowest percentage in the Slow bin of the ToxCast test set.



Fig S1*.*1 Histogram of clearance values (µL/min/106 cells) for all *in vitro* values (ChEMBL and ToxCast), colored according to the classification as very slow , slow, fast or very fast. All values have been corrected for *in vitro* binding per assay, as described in S1.1.2.

S1.2 Data Collection and Curation for fup

 *In* vitrovalues for *fup* were assembled starting from the dataset of Ingle et al.2, a collection of 1651 (1245 pharmaceuticals and 406 ToxCast) compounds curated from the literature 3-6 In contrast to Ingle et al.2, who predominantly used pharmaceutical compounds for training, we sought to both train and test models over a broad chemical domain. Thus, this dataset was augmented and re-assembled in the following way. First, a set of 200 pharmaceuticals was withheld as a test set. The remaining 1451 chemicals were augmented with 317 new compounds from the more recent Wambaugh et al.7. Of these 317 new chemicals, 70 could be considered pharmaceutical-like based on comparisons with DrugBank8 ( a database of pharmaceuticals from the University of Alberta), and 247 were likely non-pharmaceuticals. The resulting 1768 chemicals consisted of 1115 pharmaceuticals and 650 ToxCast chemicals. In order to create a final training set in which the distributions of *fup* were similar between pharmaceuticals and ToxCast compounds, 655 pharmaceuticals were randomly selected using the R package *caret*9, 10. Combining these with the 650 chemicals from ToxCast resulted in a total training set of 1310 compounds. Meanwhile, the remaining pharmaceuticals (458) were employed as a second drug test set. Finally, we used the same 99 ToxCast chemicals used to evaluate the *Clint* model described above as a third test set. Note that the above number of chemicals was later reduced for several sets due to incalculable PaDEL descriptors. See Main Text 2.2.3 for exact details.

S1.3. Random Forest Classification Modeling of *Clint*

 Random Forest (RF) Classification was used to construct a QSAR model of binned values of *Clint* using the bins described above, and the descriptor sets described below. While the descriptors are used as predictors in models, we will subsequently continue to refer to them as descriptors throughout for continuity. The three general steps of the model-building process included 1) assembling descriptor sets to construct models, 2) model construction, and 3) the determination of a source-specific optimal model. Chemical structures were expressed as 2D Simplified Molecular Input Line Entry System (SMILES) strings, which were obtained for each chemical based on the EPA Distributed Structure-Searchable Toxicity Database (DSSTox) substance identifier (DTXSID) or ChEMBL ID. Chemical structures were prepared for QSAR development by removing salts and neutralizing charges. Four independent open-source descriptor sets were utilized: PaDEL,21 OPERA,22 ToxPrints, 23, 24 and MACCS.25, 26 Of these descriptors sets, PaDEL, ToxPrints, and MACCS are different systems of describing chemicals through chemical structure fingerprints. In contrast, OPERA represents a collection of physicochemical properties predicted by CDK-structural and PaDEL fingerprints.27 Descriptor sets for chemicals were calculated based on their 2D structures and pruned to remove uninformative low variance (<5%) and strongly correlated (>95%) descriptors. For *Clint*, the final datasets included 710 PaDEL descriptors, 17 OPERA descriptors, 79 ToxPrints, 116 MACCS fingerprints, and a combined set of all descriptor sets with 917 descriptors referred to hereafter as the “All Datasets” set. Note, the pruning step was carried out for each source, so the 917 descriptors in the All Datasets set includes some excluded from the individual source-specific sets. Lastly, all descriptors in each descriptor set were centered and scaled to 1 standard deviation with respect to the training set.

 To prevent overfitting while maintaining generalizability, a recursive feature elimination approach was used to find the minimum set of descriptors necessary to maintain predictive ability. First, a model was fit for each descriptor set (PaDEL, OPERA, ToxPrints, MACCS, or All Datasets), with all descriptors ranked based on importance. Importance is a measure that determines how much the out-of-bag accuracy decreases when each descriptor is in turn left out of the model.28 Using the rank-ordered descriptors per source when all descriptors were included, a series of models was constructed with increasing numbers of descriptors according to the following sequence: 1,2,3,..10, 20, 30..,100, 200, 300..etc, with the exact number of models varying depending upon the number of descriptors available. Following model construction in one step of a series, importance was recalculated and used to select the descriptors for the next step. At each step of the series, a 5-fold cross-validation was conducted. In this procedure, the training set was randomly split into 5 equal subsets, and a series of 5 models were iteratively constructed using 4 of the subsets and evaluated against the remaining set. Each subset contained equal representation from each of the classification bins. In addition to the prescribed series progression above, all model series also included both a model with only 1 descriptor, and a model with the total number of descriptors for the descriptor source. All models were created using the R package *randomForest.*25, 29

 In RF models, an important hyperparameter is the tuning parameter (*mtry*), which corresponds to the number of descriptors to be considered for each split in the boot-strapped construction of classification trees29. This parameter was tuned by iteratively training models with increasing numbers of descriptors (starting with lowest integer of the square root of the number of descriptors) and calculating the mean absolute error (MAE) each time. Then, *mtry* is set to the number of variables that resulted in the absolute lowest MAE. This process was carried out at each step of each series of model fits. Finally, all models were constructed with 500 trees, as preliminary work found that increasing the number of trees in models to 5000 trees did not significantly alter model accuracy.

 Following model construction, model accuracy and standard errors at each step of each recursive feature elimination sequence described above were calculated by averaging the accuracy (via the R package *caret9, 10*) of the 5 cross-validation models involved. In addition, as the assay based *Clint* values in the fast and very fast categories may exceed physiological limits of liver blood flow30, a parallel assessment of classification models with a combined fast and very fast bin (i.e., 3 bins in total) was also undertaken.

 To find the most accurate and parsimonious model for each descriptor source and bin (i.e., the source/bin-specific optimal model), the model subset with the highest average accuracy (source/bin-specific best model) was first identified. This was generally the set with the largest number of descriptors. The source-specific optimal model was then defined as the model with the least number of descriptors that had an average accuracy within 1 standard error of accuracy of the source-specific best model. See the Supplemental Information (S2.6) for a graphical representation of the selection of source-specific optimal models. In these figures, the optimal source-specific model for both 3- and 4-bin models are marked with vertical red and blue vertical lines.

 Finally, the accuracy of each source/bin-specific optimal model was assessed against the assembled test sets, including the ToxCast set and the more pharmaceutical-centric ChEMBL set. The statistical significance of the accuracy of prediction was assessed against the no-information rate, which represents the likelihood of success if classes were assigned by random chance. In this case, the no-information rate is simply the proportion of the largest class. Lastly, out of all the source/bin-specific optimal models, the one with the highest average accuracy was defined as the best overall model. This best overall model was utilized for all downstream QSAR-based *Clint* predictions.

S1.4. Random Forest Regression Modeling of *fup*

 Random forest regression was used to model fraction unbound in plasma (*f*up) using the *caret9, 10* package of program R. The hyperparameter *mtry* was set at a constant of 1/3 number of descriptors, and all models were fit with 500 trees. Performance statistics to evaluate models included the root mean square error (RMSE) and mean absolute error (MAE):

RMSE = (1)

MAE = (2)

Prior to model selection, four different transformations of *in vitro* *fup* values were evaluated to determine the optimal transformation:

 (4)

To evaluate these transformations, a recursive feature elimination (RFE) process was carried out with 5-fold cross validation for each transformation(i) using the “All Features” set. In this process, predictions (ypred,i) and root mean square error (RMSE\*) metrics were calculated and averaged over the 5-folds in the transformation space. We note this as the RMSE\* to contrast it from untransformed RMSE which was used later to assess model performance. For comparison, RMSE\* was normalized by sigma, the standard deviation of the transformed ypred,i (RMSE\*/sigma). The optimal transform was determined by the overall lowest normalized RMSE\* (see Supplemental Information (S2.14)).

 For model selection we carried out another RFE with 5-fold cross validation with each descriptor set using the optimal transformation (square root). The general details of the RFE process were similar to those described in S1.3 except that descriptors were not re-ordered at each step, 5 fold-averaged RMSE\* was used as the evaluation metric, and the 95% confidence interval was used as a measure of uncertainty around averaged values. The selection of source-specific optimal models is graphically show in the Supplemental Information (S2.9). In these figures, the red dashed lines (intersecting y-axis) represent an empirical asymptote on RMSE\*, calculated based on the mean RMSE\* of the models. They are the upper 95% confidence intervals on the mean RMSE\* of the three models with most descriptors, horizontally projected. The blue dashed line (intersecting x-axis) represents the point where the recursive feature elimination has arrived at the minimum number of descriptors for which additional descriptors make no improvement on model fit, based on RMSE\*. The source-specific optimal models were chosen as the model sets at this point.

Following model selection, model predictions were transformed back to the original space by the following prior to evaluation against test sets:

 (7)

S1.5 Bioactivity-Exposure ratio (BER) and data sources

*S1.5.1. Bioactivity Data and OED conversion*

 *In vitro* toxicity data were from the ToxCast and Tox21 programs, multi-agency collaborative programs aimed at characterizing the toxicity chemicals using *in vitro* methods31. These data were concentrations eliciting a half-maximal response (that is AC50 values) in a large number and variety (>1470 assays) of *in vitro* screening assays for bioactivity. Both ToxCast and Tox21 test chemicals in concentration-response across panels of *in vitro* assays (>1100 and >50 assays, respectively). Most chemical-assay combinations do not show a systematic relationship between concentration and response. An AC50 is calculated only if a concentration-dependent response is observed, a so-called “hit” or active assay. Data were available for 93 chemicals of the ToxCast test set used in the above QSARs, with an average of about 90 active assays (that is, datapoints) per chemical. For the 7564 chemicals included in the Tox21 analysis, there was an average of about 36 active assays per chemical.

 After the assembly of data, the bioactive dose (Q10) was calculated as the lowest 10th percentile (Q10) of the half-maximal activity concentrations (AC50) across the ToxCast screening assays. From this, an OED for the Q10 was determined with the 3 compartment steady state model of the *httk* R package32 which predicted distributions of steady state concentrations to account for variability in toxicokinetic and physiological parameters across individuals. From these distributions, the upper 95th percentile served as an approximation for sensitive population internal tissue dose upon exposure. From these concentrations, OED concentrations (in mg/kg/day) were calculated using the following equation15:

 (8).

*S1.5.2 Expected Exposure Data*

Expected exposure estimates for chemicals in terms of OED (in mg/kg bodyweight/day) were generated in two ways. First, for chemical for which urine analytes were available (ToxCast:10 Tox21:99) in the National Health and Nutrition Examination Survey (NHANES)33, OED values were inferred using a heuristics methodology34. For the remaining chemicals (ToxCast:83, Tox21:7471), OED values were predicted using the Systematic Empirical Evaluation of Models (SEEM) framework35. The SEEM framework is a meta-modelling approach that combines predictions from 13 different exposure models, each parameterized to predict hundreds to thousands of chemicals, into a weighted Bayesian analysis to produce consensus distributions of predicted exposure. The SEEM approach has the benefit of including exposure estimates from both near and far-field sources, and of providing a quantitative estimate of uncertainty.

REFERENCES

1. Kilford, P. J.; Gertz, M.; Houston, J. B.; Galetin, A., Hepatocellular binding of drugs: correction for unbound fraction in hepatocyte incubations using microsomal binding or drug lipophilicity data. *Drug Metab Dispos* **2008,** *36*, (7), 1194-7.

2. Ingle, B. L.; Veber, B. C.; Nichols, J. W.; Tornero-Velez, R., Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability. *J Chem Inf Model* **2016,** *56*, (11), 2243-2252.

3. Obach, R. S.; Lombardo, F.; Waters, N. J., Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug Metab. Dispos.* **2008,** *36*, (7), 1385-1405.

4. Zhu, X. W.; Sedykh, A.; Zhu, H.; Liu, S. S.; Tropsha, A., The Use of Pseudo-Equilibrium Constant Affords Improved QSAR Models of Human Plasma Protein Binding. *Pharm Res-Dordr* **2013,** *30*, (7), 1790-1798.

5. 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. *Toxicol Sci* **2012,** *125*, (1), 157-74.

6. 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. *Toxicol Sci* **2015,** *148*, (1), 121-136.

7. 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. In 2019.

8. Wishart DS, F. Y., Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M., DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **2017**.

9. Kuhn, M., Caret: Classification and regression training. R package version 6.0-85. In R Project, <https://CRAN>. R-project. org/package= caret: 2020.

10. Kuhn, M. *caret: Classification and Regression Training*, 2019.

11. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P., The ChEMBL bioactivity database: an update. *Nucleic Acids Res* **2014,** *42*, (Database issue), D1083-90.

12. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **2012,** *40*, (Database issue), D1100-7.

13. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R., The ChEMBL database in 2017. *Nucleic Acids Res* **2017,** *45*, (D1), D945-D954.

14. 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. *Chem. Res. Toxicol.* **2016,** *29*, (8), 1225-1251.

15. 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. *Toxicol Sci* **2010,** *117*, (2), 348-358.

16. Wetmore, B. A.; Wambaugh, J. F.; Ferguson, S. S.; Sochaski, M. A.; Rotroff, D. M.; Freeman, K.; Clewell, H. J.; 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. *Toxicol Sci* **2012,** *125*, (1), 157-174.

17. 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 Metab Dispos* **2002,** *30*, (8), 892-6.

18. Pearce, R. G.; Setzer, R. W.; Strope, C. L.; Wambaugh, J. F.; Sipes, N. S., httk: R Package for High-Throughput Toxicokinetics. *J Stat Softw* **2017,** *79*, (4), 1-26.

19. Ito, K.; Houston, J. B., Comparison of the Use of Liver Models for Predicting Drug Clearance Using in Vitro Kinetic Data from Hepatic Microsomes and Isolated Hepatocytes. *Pharm Res-Dordr* **2004,** *21*, (5), 785-792.

20. Brown, R. P.; Delp, M. D.; Lindstedt, S. L.; Rhomberg, L. R.; Beliles, R. P., Physiological parameter values for physiologically based pharmacokinetic models. *Toxicol Ind Health* **1997,** *13*, (4), 407-84.

21. Yap, C. W., PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* **2011,** *32*, (7), 1466-74.

22. Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J., OPERA models for predicting physicochemical properties and environmental fate endpoints. *J Cheminform* **2018,** *10*, (1), 10.

23. Wang, J.; Hallinger, D. R.; Murr, A. S.; Buckalew, A. R.; Lougee, R. R.; Richard, A. M.; Laws, S. C.; Stoker, T. E., High-throughput screening and chemotype-enrichment analysis of ToxCast phase II chemicals evaluated for human sodium-iodide symporter (NIS) inhibition. *Environ. Int.* **2019,** *126*, 377-386.

24. Yang, C.; Tarkhov, A.; Marusczyk, J.; Bienfait, B.; Gasteiger, J.; Kleinoeder, T.; Magdziarz, T.; Sacher, O.; Schwab, C. H.; Schwoebel, J.; Terfloth, L.; Arvidson, K.; Richard, A.; Worth, A.; Rathman, J., New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. *J Chem Inf Model* **2015,** *55*, (3), 510-28.

25. Guha, R., Chemical Informatics Functionality in R. *Journal of Statistical Software* **2007,** *6*, (18).

26. MDL Information Systems Inc. *MACCS keys.* , 14600 Catalina Street, San Leandro, CA 94577.

27. Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Cherto, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C., The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* **2017,** *9*, (1), 33.

28. Breiman, L., Random Forests. *Machine Learning* **2001**, 5-32.

29. Liaw, A.; Wiener, M., Classification and Regression by randomForest. *R News* **2002,** *2*, (3), 18-22.

30. Obach, R. S.; Baxter, J. G.; Liston, T. E.; Silber, B. M.; Jones, B. C.; MacIntyre, F.; Rance, D. J.; Wastall, P., The prediction of human pharmacokinetic parameters from preclinical and in vitro metabolism data. *J. Pharmacol. Exp. Ther.* **1997,** *283*, (1), 46-58.

31. U.S. EPA ToxCast & Tox21 Data Spreadsheet for In-vitro DBv3.2. <https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data> (02/10/2020),

32. 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* **2017,** *79*, (4).

33. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), National Health and Nutrition Examination Survey Data. In U.S. Department of Health and Human Services, Centers for Disease Control and Prevention: Hyattsville, MD., 2009-2010.

34. Wambaugh, J. F.; Wang, A.; Dionisio, K. L.; Frame, A.; Egeghy, P.; Judson, R.; Setzer, R. W., High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals. *Environ. Sci. Technol.* **2014,** *48*, (21), 12760-12767.

35. Ring, C. L.; Arnot, J. A.; Bennett, D. H.; Egeghy, P. P.; Fantke, P.; Huang, L.; Isaacs, K. K.; Jolliet, O.; Phillips, K. A.; Price, P. S.; Shin, H. M.; Westgate, J. N.; Setzer, R. W.; Wambaugh, J. F., Consensus Modeling of Median Chemical Intake for the U.S. Population Based on Predictions of Exposure Pathways. *Environ Sci Technol* **2019,** *53*, (2), 719-732.