# Defining uncertainty in publicly available high-throughput screening data from the ToxCast program


**Jason Brown[1,3], Eric D. Watt[2], R. Woodrow Setzer[3], Richard Judson[3], Katie Paul Friedman[3]**

[1]Oak Ridge Institute for Science Education Student Contractor, Oak Ridge, TN, 37831; [2]Pfizer, New York City, Pharmacokinetics, Dynamics and Metabolism Research Unit
[3]U.S. Environmental Protection Agency, National Center for Computational Toxicology, Research Triangle Park NC USA



**3368/P151**
March 11-15, 2018
Society of Toxicology Annual Meeting San Antonio, TX



Jason Brown | brown.jason@epa.gov | 919-541-5098


www.epa.gov

*This poster does not necessarily reflect EPA policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.*

## Abstract

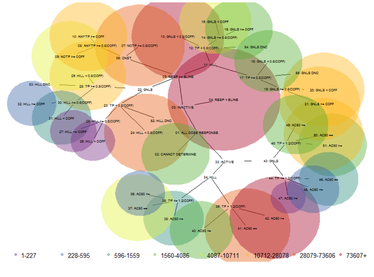

The US Environmental Protection Agency (EPA) ToxCast data pipeline (tcpl [1]) is currently applied to >1000 assay endpoints to enable first tier data processing of heterogeneous data types obtained from high-throughput screening for bioactivity. One result of this initial analysis is generation of activity concentration information, including the 50% activity concentration (AC50) for each chemical sample-assay endpoint pair for which the data can be fit to a curve. As with any curve-fitting exercise, there are multiple sources of potential variability in these AC50 values, resulting from biological variance, experimental error, or curve-fitting procedures. The primary objectives of this work were to: (1) implement a recently developed bootstrap resampling method, available as the toxboot R package, for generation of uncertainty information related to the AC50 values generated using tcpl; (2) to develop a new level of tcpl to house a summary of these uncertainty information to inform filtering or modeling of ToxCast data; and, (3) derive general trends in the uncertainty information to develop a greater understanding of the data currently in the ToxCast database, invitrodb. Briefly, toxboot uses a smooth nonparametric bootstrap resampling to add random normally distributed noise to give a resampled set of concentration-response values. The resampled data is fit to the three ToxCast models and repeated 1000 times before it is stored in a Mongo database including over 50 variables relating to the model fitting parameters. The resulting data were used to generate point estimates, winning model, and hit call for each of the 1000 resamples. Various summary statistics such as hit percent, median AC50, AC50 confidence interval, etc. were generated based on the toxboot resampling. Hit percent, i.e. the probability that based on the set of resampled data there would be a positive hit-call, may be a useful value for predictive modeling in place of binary hit-call; 78% of the positive hit-calls in invitrodb corresponded to a hit percentage ≥ 90 and the median AC50 confidence interval width was 0.368 log10 micromolar units. The AC50 median and confidence interval may be useful values for quantifying AC50 variability. Application of toxboot to ToxCast data provides a statistically robust means of estimating the uncertainty in AC50 values and evaluating the reproducibility of curve fits from the ToxCast pipeline. *This abstract does not necessarily reflect U.S. EPA policy.*


## Current ToxCast Pipeline (tcpl) curve classification

### TCPL level 5: Fit Categories

- Goal: categorize curves into the fit categories (fitc) to enable faster quality control checking and identification of potential false positives.
- Some categories are more likely to have a higher level of uncertainty associated with the curve fit and/or the hit-call.

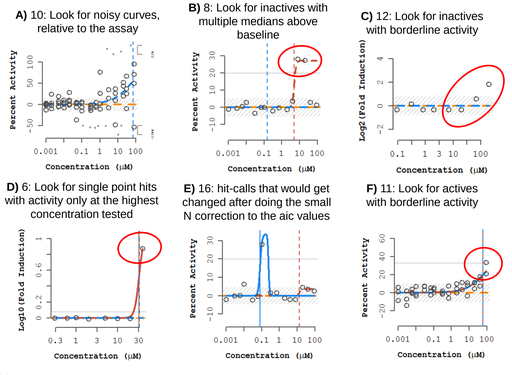
**Figure 1: Fitc relative distribution in invitrodb_v2.**




### TCPL level 6: Caution Flags

- Goal: further identify series that have potential for false positive or false negative results.

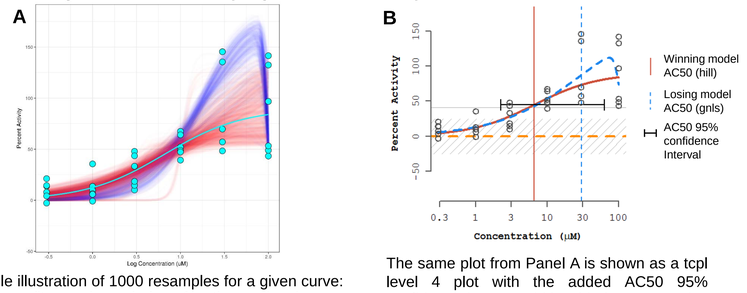
**Figure 2A-F: Examples of specific caution flags.**




## Toxboot builds uncertainty quantification into the ToxCast Pipeline

- Toxboot (R package available on CRAN [2]) uses smooth nonparametric bootstrapping, a statistical method that uses resampling and added noise (mean zero, standard deviation equal to the median absolute deviation of the response at the lowest concentrations) to determine uncertainty in a series.
- As hit-calls are binary (positive or negative), they are susceptible to variability and uncertainty in curve-fitting.
- If following resampling with added random, normally-distributed noise to the series, similar curve-fits and hit-calls are produced, one could be more confident in the results.

**Figure 3A and B: resampling illustration and implementation of toxboot data**

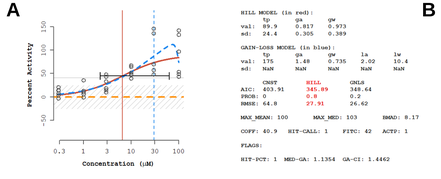



Example illustration of 1000 resamples for a given curve: blue curve fits used a gain-loss function and red curve fits used a Hill fit (from tcpl).

The same plot from Panel A is shown as a tcpl level 4 plot with the added AC50 95% confidence interval width added to summarize the toxboot uncertainty estimation.

## Toxboot implementation for 2.2 million curves

**Figure 4A and B: Example illustrations of toxboot results.**




B

```
HILL MODEL (in red):
          tp      ga      gw
val: 89.9     0.817   0.973
sd:  24.4     0.305   0.389

GAIN-LOSS MODEL (in blue):
          tp      ga      gw      la      lw
val: 175      1.48    0.735   2.02    10.4
sd:  NaN      NaN     NaN     NaN     NaN

          CNST    HILL    GNLS
AIC:  403.91    345.89    348.44
BMDL:  0        0.8       0.2
RMSE:  44.8     27.91     26.42

MAX_MEAN: 100        MAX_MED: 103        RMAD: 8.17

COFF: 40.9   HIT-CALL: 1    FITC: 42   ACTP: 1

FLAGS:

HIT-PCT: 1  MED-GA: 1.1354  GA-CI: 1.4462
```

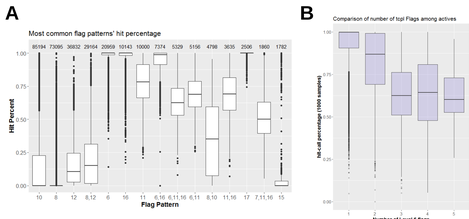
| Stored Parameter | Description |
|---|---|
| Hit_pct | Hit Percentage |
| Modl_ga_min, Modl_ga_max, Modl_ga_delta | Lower, upper, and width of the AC50 confidence interval |
| Modl_ga_med | Median AC50 calculated from bootstrapping |
| Modl_gw_med | Median hill coefficient calculated from bootstrapping |

- **Challenge 1: Computational time.** With 2.2 million concentration response series in invitrodb_v2, it would take years on a single core machine to process 1000 resamples per curve.
- **Solution 1: Parallel processing.** By scaling the processing up to run on a server with ~200 cores, we could reduce the amount of time to bootstrap the entire set of data to < 3 weeks.
- **Challenge 2: Data size.** For 2.2. million curves in invitrodb_v2, Toxboot results are ~ 1 Terabyte in size.
- **Solution 2: Use a NoSQL type database such as MongoDB**.
- **Challenge 3: Key parameters to store.** Each of the resampled series could be processed similarly to the level 5 processing done in tcpl. This includes determining the wining model, hit-call determination, calculating point-of-departure estimates, and fit category selection.
- **Solution 3: Separate database resources.** All resampled data are stored in MongoDB, and summary parameters (Figure 4B) are stored back to a new level 7 table in invitrodb (pre-release).

## Hit Percent: A non-binary hit-call determination

- Hit percent = Number of positive hit-calls out of the total number of resamples.
  - o Level 5 tcpl processing was completed for each of the 1000 resamples curve
  - o Hit-call can be determined for each of the resamples.
- Potentially more informative because the binary hit-call is susceptible to both false positive and false negative results.
- 78% of the positive hit-calls in invitrodb corresponded to a hit percentage ≥ 90.

**Figure 5A and B: Hit percent vs. caution flags**



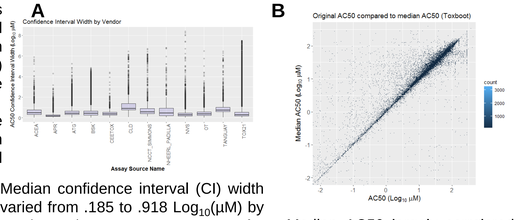

These 15 flag patterns cover over 95% of the different types of flag patterns in ToxCast.

Curves with multiple flags have a wide range of hit percentages

## Median AC50 Confidence Interval

- From the same 1000 resamples that the hit percent was generated from, the level 5 tcpl processing can generate an AC50 (or AC10, AC20, AUC etc.) and generate a 95% confidence interval.
- Median AC50 value can be calculated, and the combination of these statistics may be useful for quantifying AC50 variability.

**Figure 6A and B: Median AC50 CI and reproducibility of AC50**




Median confidence interval (CI) width varied from .185 to .918 $Log_{10}$(μM) by vendor (as a surrogate for technology). Larger CI width corresponds to less certain fits.

Median AC50 largely correlated with AC50 from tcpl, but with notable outliers.

## Conclusions

- The advantage of Toxboot is the ability to use statistical methods to better characterize the uncertainty and reproducibility of curves.
- Continuous variables such as hit percentage, and an ac50 confidence interval, allow for more fine-grained examination of the certainty in curve-fitting.
- By implementing Toxboot across all curves in the publicly available high-throughput screening data from the ToxCast program, an additional resource for uncertainty quantification is available for others to use.
- In the future, Toxboot summary statistics, fit categories, and flags could be used as features in a machine learning model used to predict goodness of fit for curves in the database.


References
[1] Filer, D.L., et al., *tcpl: the ToxCast pipeline for high-throughput screening data.* Bioinformatics, 2016. **33**(4): p. 618-620.
[2] Watt et al. (submitted). *Uncertainty Quantification in ToxCast High Throughput Screening.* 2017.