ToxCast Pipeline, Example, and Building Additional Context for Use

Katie Paul Friedman, PhD

National Center for Computational Toxicology, Office of Research and Development, US EPA

Presentation for the SETAC NA Focused Topic Meeting: High-throughput screening and environmental risk assessment

April 17, 2018

Overview: a talk in 3 parts

- Part I: Brief overview of the ToxCast Data Pipeline (tcpl).
- Part II: Example of using both tcpl and external analysis for the CEETOX high-throughput H295R (HT-H295R) steroidogenesis assay.
- Part III: Adding context for use of ToxCast data: exploring uncertainty in ToxCast.

Part I: Overview of ToxCast and the ToxCast Pipeline

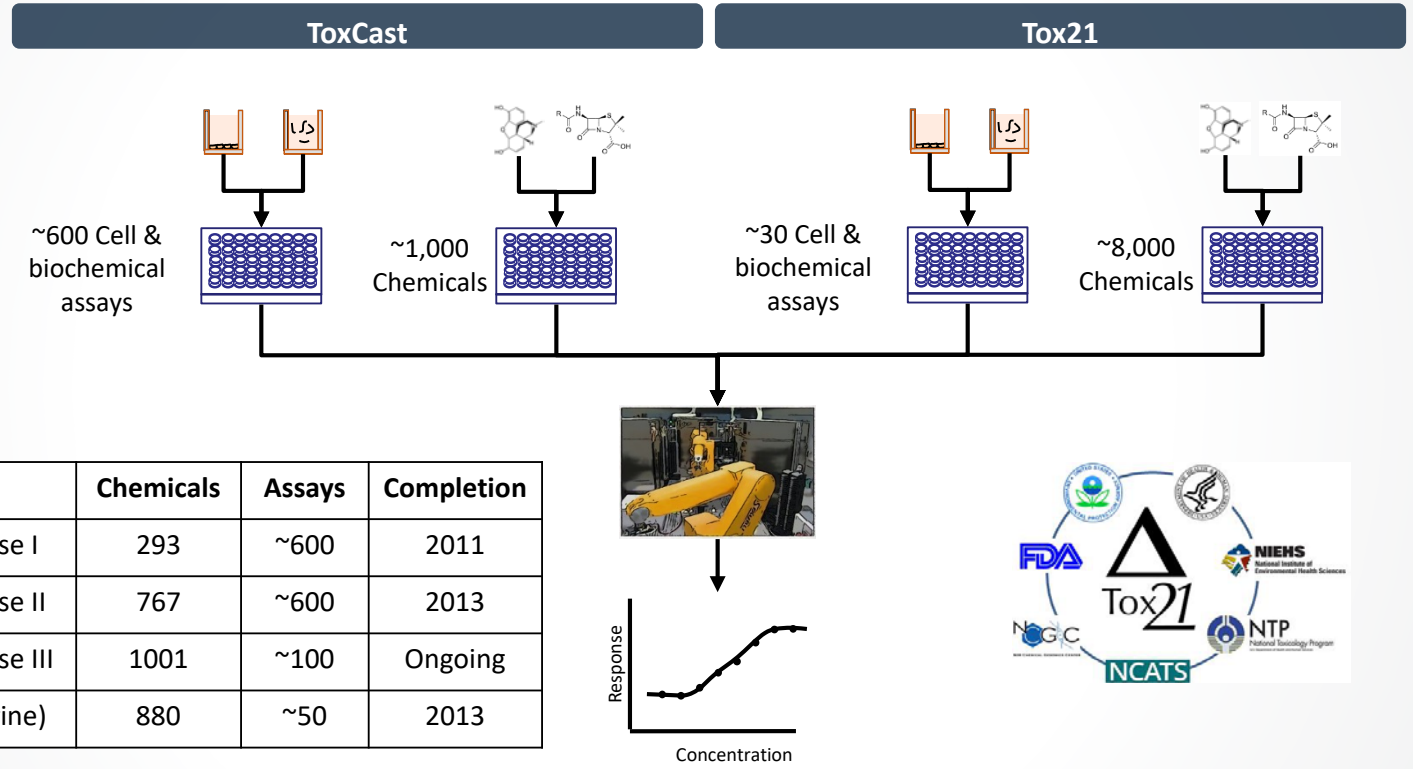
ToxCast Dashboard (current most-detailed assay information interface): https://actor.epa.gov/dashboard/

CompTox Dashboard (many data streams, currently centered on chemistry; Williams et al. 2017 PMID 29185060): https://comptox.epa.gov/dashboard

Data downloads (download databases and supporting data files):

https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data

High-Throughput Bioactivity Screening: ToxCast and Tox21

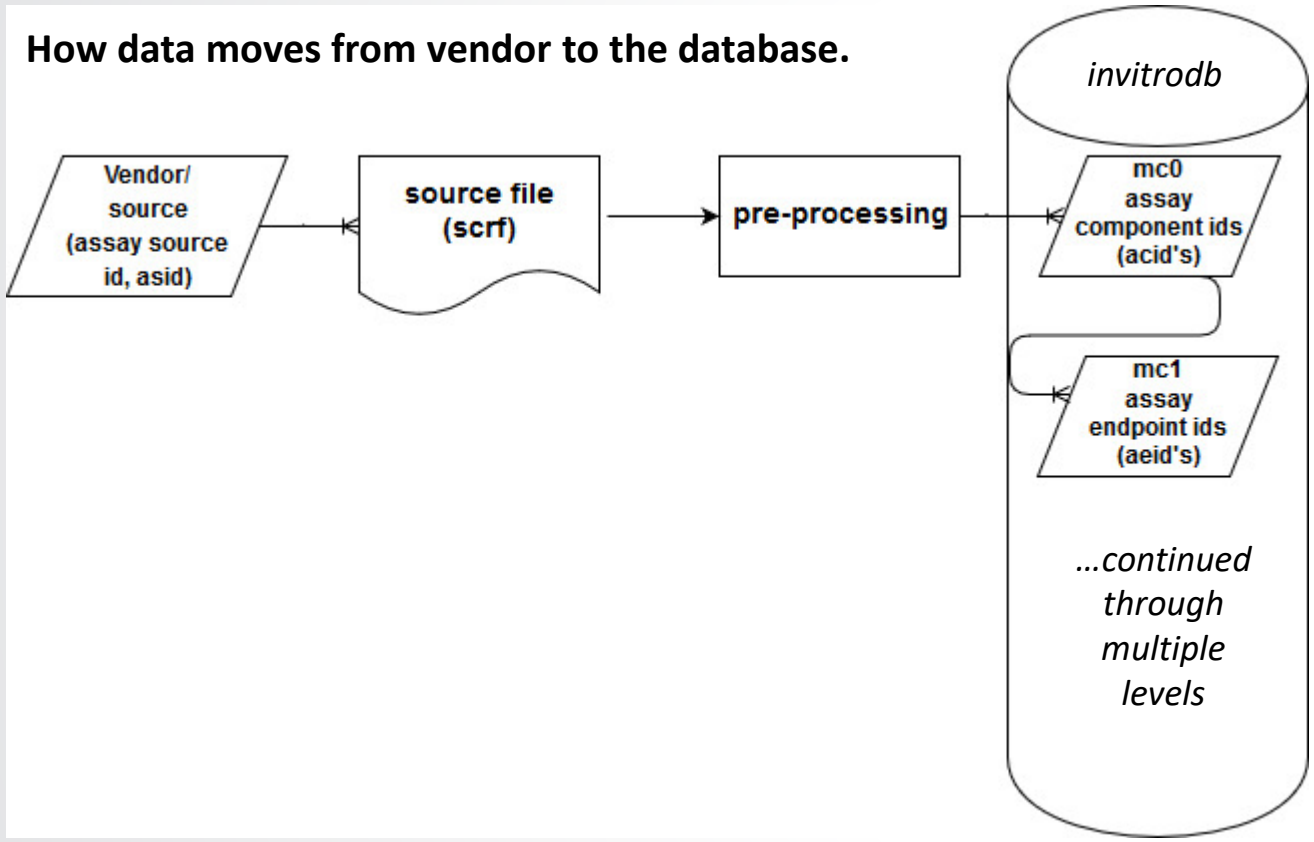


Set	Chemicals	Assays	Completion
ToxCast Phase I	293	~600	2011
ToxCast Phase II	767	~600	2013
ToxCast Phase III	1001	~100	Ongoing
E1K (endocrine)	880	~50	2013

- All Tox21 data are analyzed by multiple partners
- Tox21 data is available analyzed in the ToxCast Data Pipeline

Organization of data entering invitrodb

How data moves from vendor to the database.

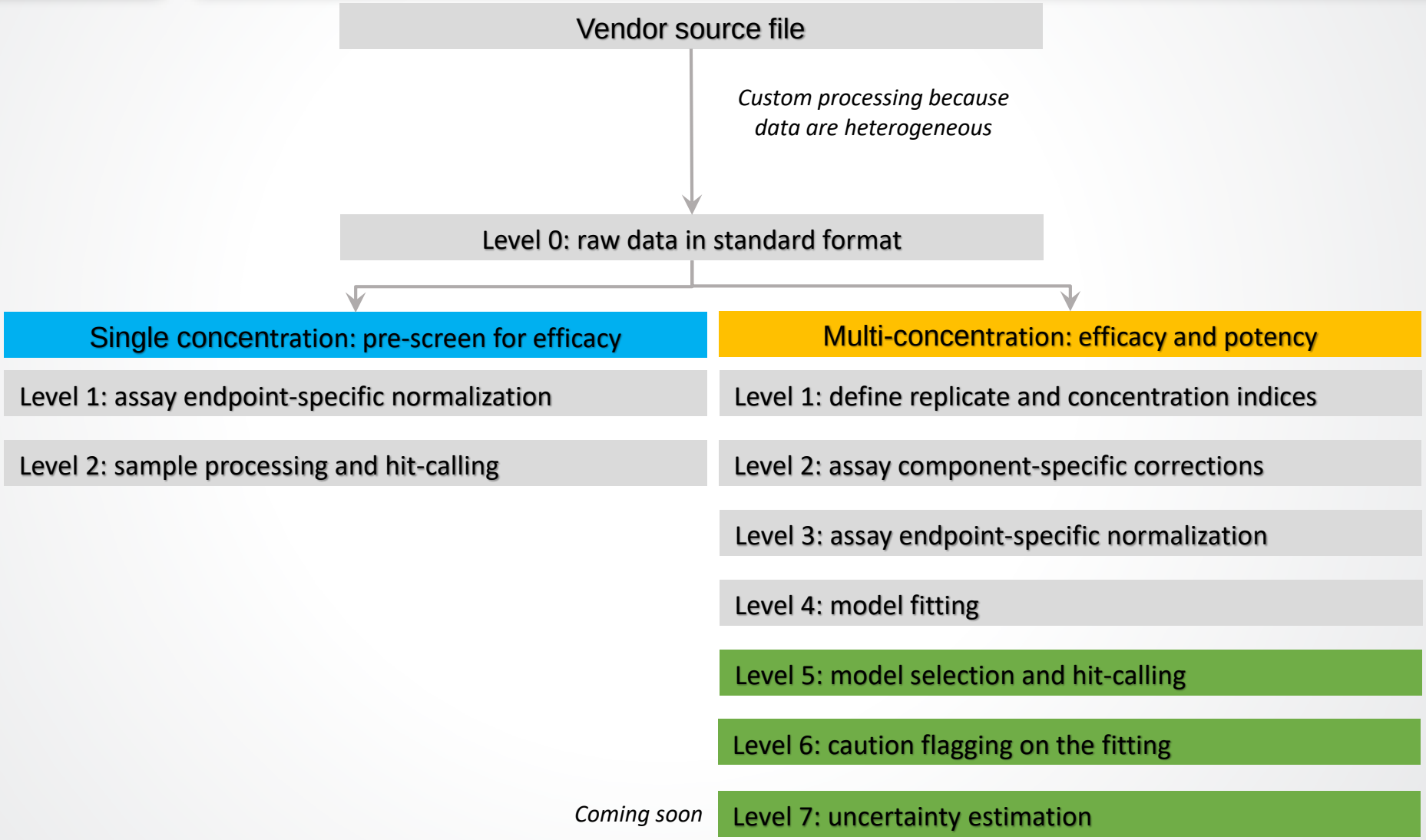


Example: asid to acid to aeid.
acid can be 1:1 or 1:many with aeid.

```
> tcplLoadAsid()
    asid           asnm
 1:    1           ACEA
 2:    2            APR
 3:    3            ATG
 4:    4            BSK
 5:    5            NVS
 6:    6             OT
 7:    7          TOX21
 8:    8         CEETOX
 9:   11            CLD
10:   12 NHEERL_PADILLA
11:   17   NCCT_SIMMONS
12:   13        TANGUAY
> tcplLoadAcid(fld='asid', val=8)
    asid acid                    acnm
 1:    8  586   CEETOX_H295R_11DCORT
 2:    8  587    CEETOX_H295R_OHPREG
 3:    8  588    CEETOX_H295R_OHPROG
 4:    8  589      CEETOX_H295R_ANDR
 5:    8  591  CEETOX_H295R_CORTISOL
 6:    8  593       CEETOX_H295R_DOC
 7:    8  594 CEETOX_H295R_ESTRADIOL
 8:    8  595   CEETOX_H295R_ESTRONE
 9:    8  597      CEETOX_H295R_PROG
10:    8  598     CEETOX_H295R_TESTO
> tcplLoadAeid(fld='acid', val=586)
   acid aeid                   aenm
1:  586  890 CEETOX_H295R_11DCORT_dn
2:  586  891 CEETOX_H295R_11DCORT_up
```

- Assay sources or vendors may send many files, which are pre-processed.
- The mc0 data in invitrodb is at the assay component level.
- At mc1, assay endpoints are defined, but it is not until normalization at mc3 that data are retrieved by assay endpoint.

Outline of the ToxCast pipeline



Part II: Example using tcpl and methods outside tcpl – high-throughput H295R (HT-H295R)

Derik Haggard, Woody Setzer, Richard Judson, and Katie Paul-Friedman

Steroidogenesis is critical for several physiological processes and modeled in the H295R cell-based assay

Steroidogenesis pathway: relevant biology

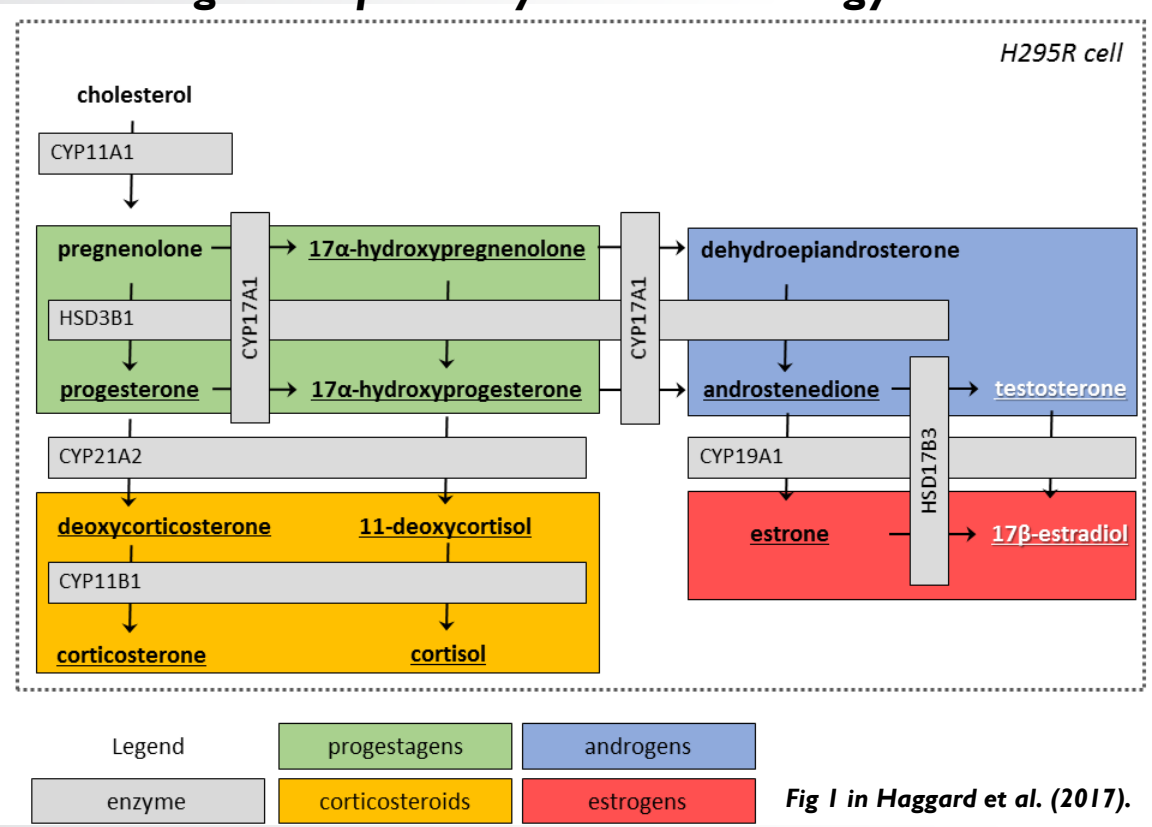
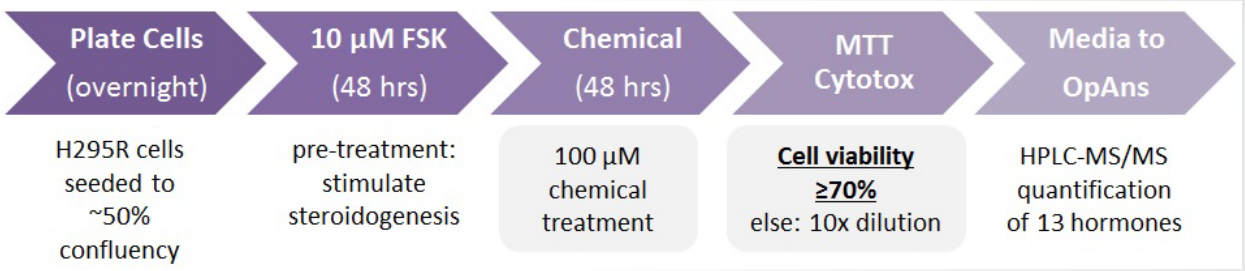


Fig 1 in Haggard et al. (2017).

High-throughput adaptation of H295R assay



- Maximized screening resource efficiency.
- 2012 unique test chemicals have been screened at a high concentration.
- # steroid hormones affected in single concentration (along with other considerations) were used to select 656 chemicals for multi-concentration screening.

Problem: How to compress 11-dimensional data to a single prioritization metric for regulators?

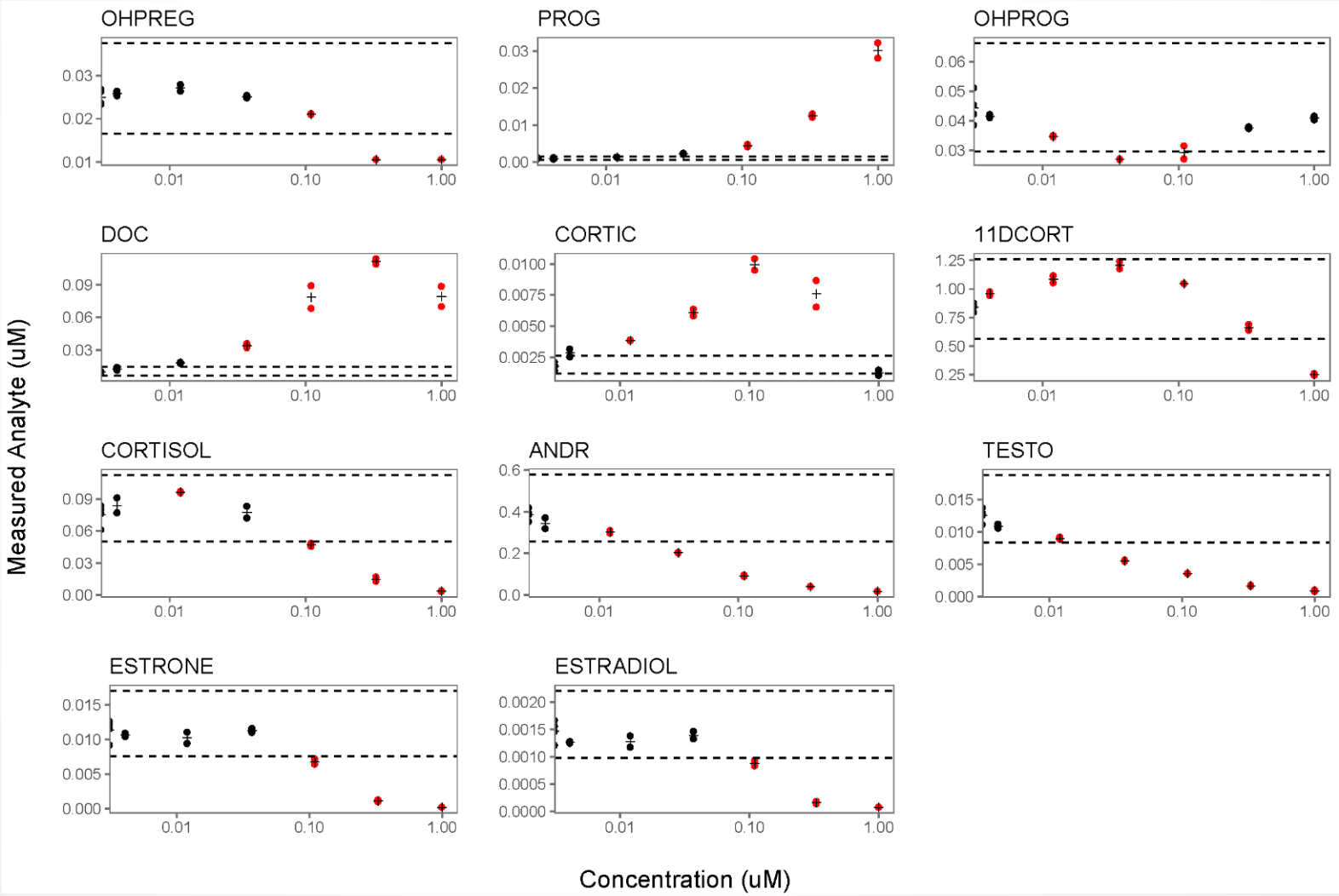
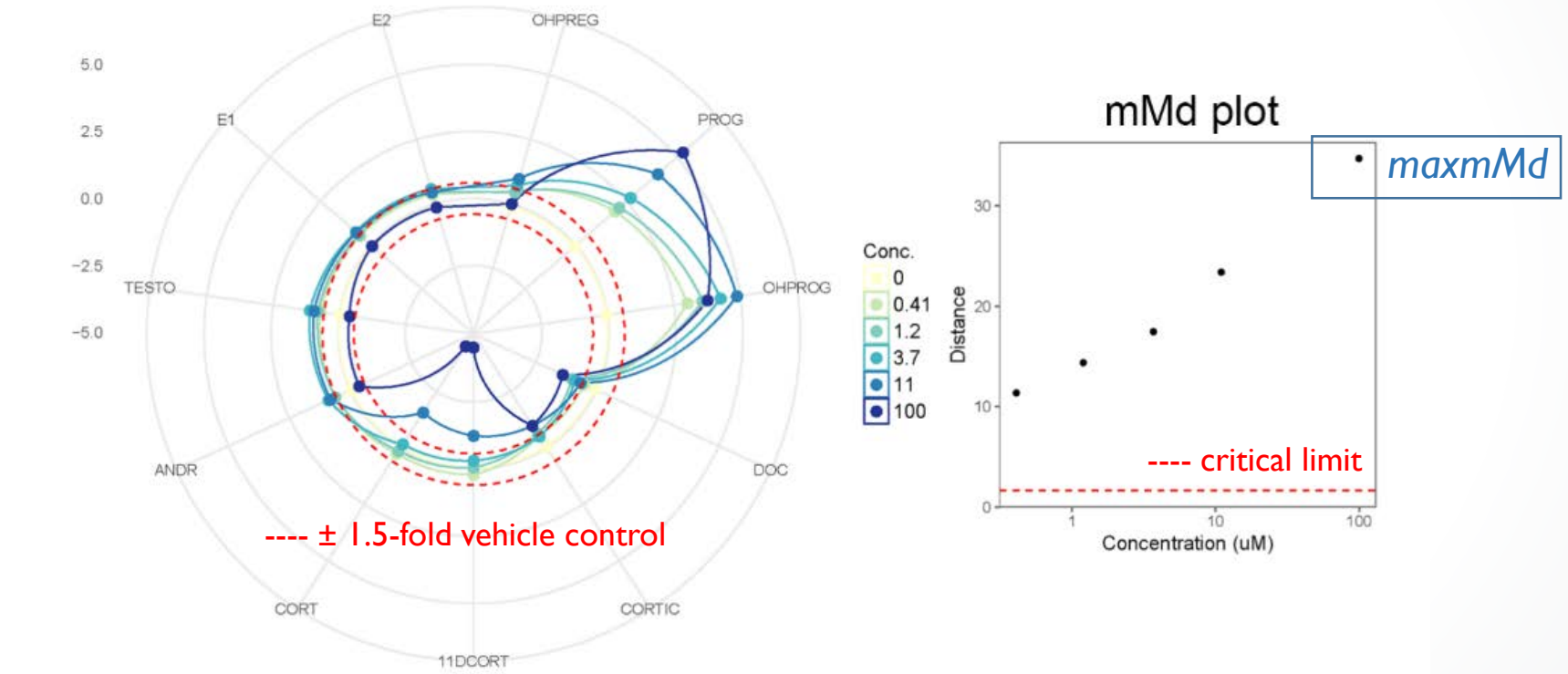


Figure 2 Haggard et al. (2018).

Using our maximum Mahalanobis distance approach to get a single prioritization metric



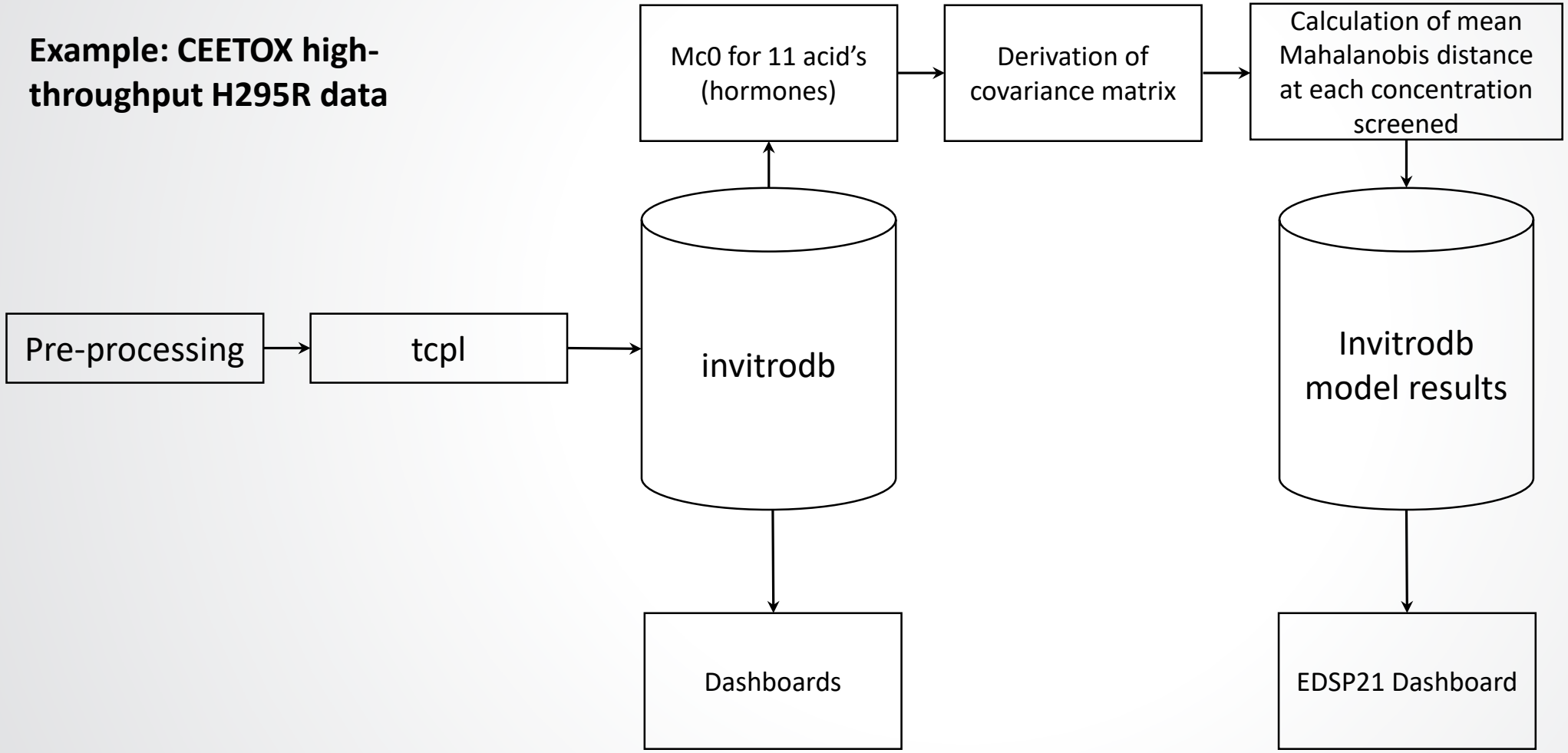
- Reduced an 11-dimensional question to a single dimension.
- Selection of the maxmMd appeared to provide a reproducible, quantitative approximation of the magnitude of effect on steroidogenesis.

Mifepristone strongly modulated progestagens with significant effects on progesterone and OH-progesterone and moderate but non-significant trends on corticosteroids and androgens, resulting in a relatively high adjusted maxmMd of 33.

Figure 5, Haggard et al. (2018).

Part II conclusions: tcpl is a first tier analysis, and some data undergo separate analysis or modeling.

Example: CEETOX high-throughput H295R data



Haggard et al. (2018) *Toxicological Sciences.* High-Throughput H295R Steroidogenesis Assay: Utility as an Alternative and a Statistical Approach to Characterize Effects on Steroidogenesis.

Also on: https://github.com/USEPA/CompTox-ToxCast-EDSPsteroidogenesis

New version coming soon

Part III: Research on uncertainty in ToxCast data

Jason Brown, Eric Watt, Woody Setzer, Richard Judson, and Katie Paul-Friedman

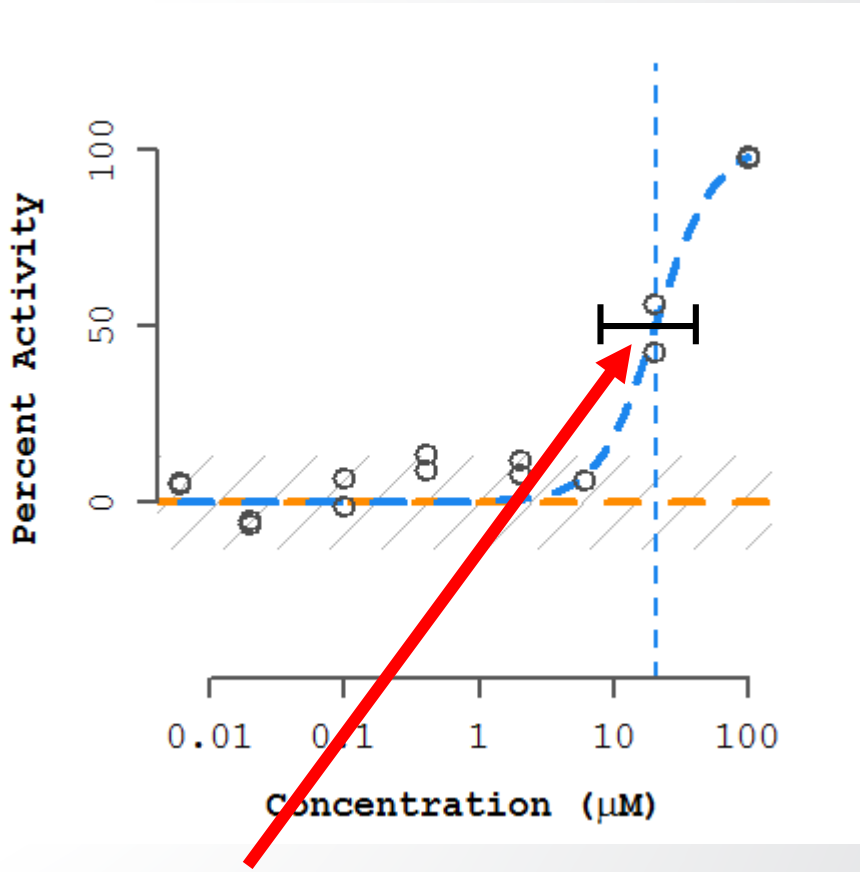
Why is defining the uncertainty in curve-fitting important?

- Appropriate conservatism in using in vitro bioactivity data as a surrogate for an *in vivo* point-of-departure.
  - Each active chemical has a distribution of AC50s.
  - The confidence interval around the lowest AC50 may produce a lower bound that is truly the most conservative value.
  - Does larger uncertainty, or a wider confidence interval for the AC50, indicate less certainty in the hitcall? Not always, but it is one important feature we could use to filter data.

- Accuracy of biological modeling: Using in vitro activity data in the development of models for specific toxicities.
  - Don’t want to include AC50 (or hitcall) from noise or overfit curves.

Defining uncertainty in curve-fitting

- Some sources of uncertainty in fitting high-throughput screening (HTS) data include:
  - Biological variance
  - Systematic error in measurement
  - Contribution of experimental design, e.g. concentration-spacing and number of concentrations
- Not quantified in tcpl currently.
- Uncertainty could be incorporated into predictive models, e.g. QSAR, hybrid descriptor sets, etc., and likely impacts predictivity of these models.
- Quantifying uncertainty may support more robust screening level risk assessment.
- Uncertainty from fitting is often conflated with uncertainty regarding the selectivity (or specificity) of a response.



How do we determine this? (Among other things)

Fit categories (fitc) follow a hierarchical tree and could potentially be used to sort curve fits.

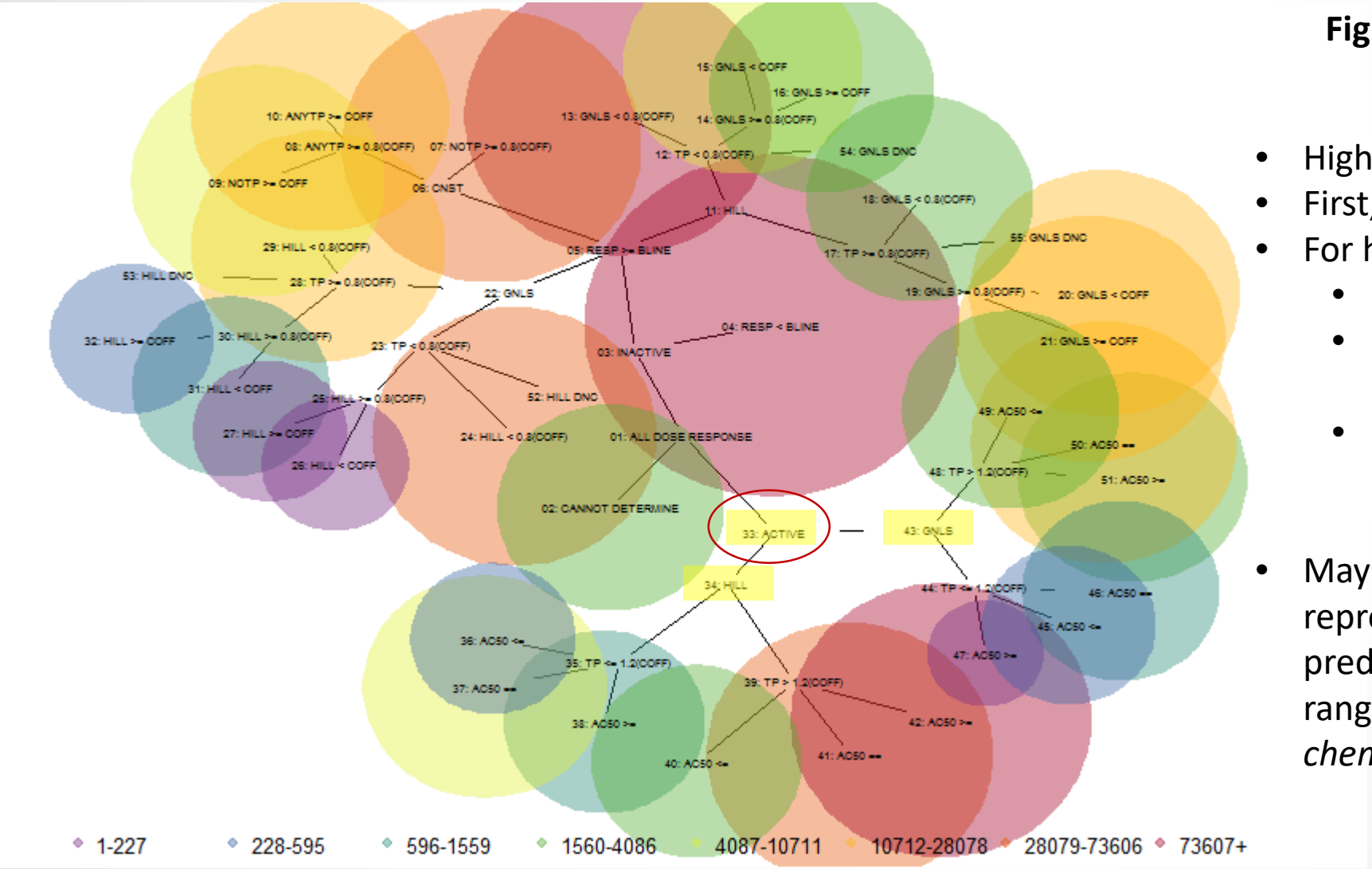
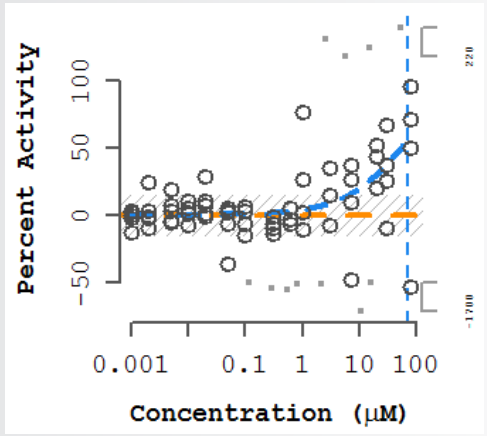


Figure 1: Relative distribution of curves by fit category in *invitrodb_v2.*

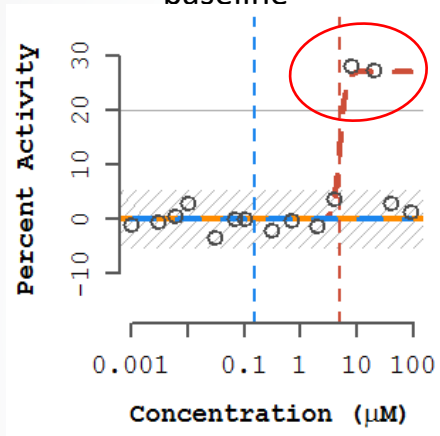
- Highest number of curves are inactive
- First, separate by hitcall (-1, 0, 1)
- For hitcall=1 [actives]:
  - separate by winning model (hill, gnls)
  - For each model, separate curves by efficacy (<1.2*coff* or ≥1.2*coff)*
  - Separate by position of AC50 with respect to the screened concentration range
- May have less confidence in the reproducibility of curves where AC50 predicted is less than the concentration range tested; *but what about reference chemicals or potent acting chemicals*?

Caution flags have also been suggested as a way to filter curves for reliability.

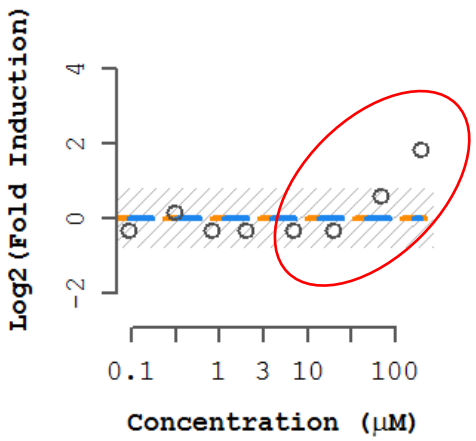
A) 10: Look for noisy curves, relative to the assay



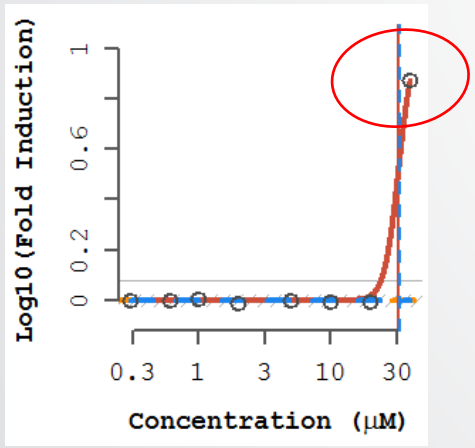
B) 8: Look for inactives with multiple medians above baseline



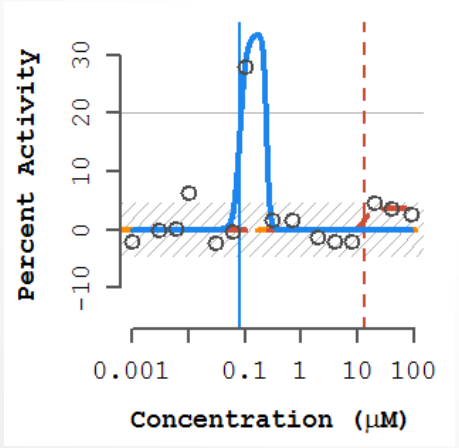
C) 12: Look for inactives with borderline activity



D) 6: Look for single point hits with activity only at the highest concentration tested



E) 16: hit-calls that would get changed after doing the small N correction to the aic values



F) 11: Look for actives with borderline activity

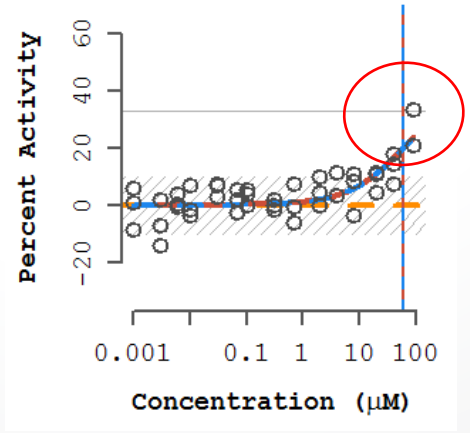


Figure 2: Curve behavior for flags associated with active curves.

- Do specific flags or numbers of flags for a specific curve fit indicate a less reliable curve fit?
- How do we benchmark the "uncertainty" in the fit to understand if flag-based filtering is only removing "poor" or "less reliable" curve fits?

State of the science: NCATS filters curves

Using Efficacy:

NCATS has used efficacy and data curve "quality"

(Huang 2016 DOI 10.1007/978-1-4939-6346-1_12 (below); Huang et al. 2014 DOI: 10.1038/srep05664)

Table 1
Amended qHTS curve classification

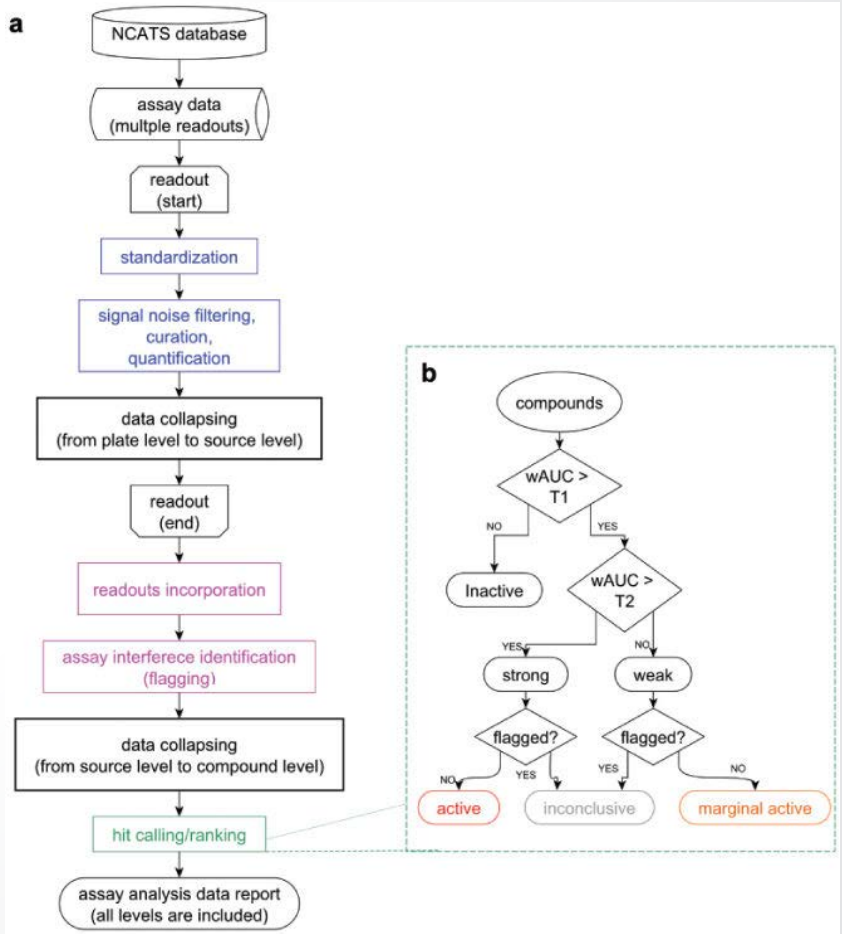
Curve class	Description	Efficacy	*p*-value[a]	Asymptotes	Inflection
1.1	Complete curve	>6SD[b]	<0.05	2	Yes
1.2	Complete curve	≤6SD; >3SD	<0.05	2	Yes
1.3	Complete curve	>6SD	≥0.05	2	Yes
1.4	Complete curve	≤6SD; >3SD	≥0.05	2	Yes
2.1	Incomplete curve	>6SD	<0.05	1	Yes
2.2	Incomplete curve	≤6SD; >3SD	<0.05	1	Yes
2.3	Incomplete curve	>6SD	≥0.05	1	Yes
2.4	Incomplete curve	≤6SD; >3SD	≥0.05	1	Yes
3	Single point activity	>3SD	NA	1	No
4	Inactive	≤3SD	≥0.05	0	No
5[c]	Inconclusive	NA	NA	NA	NA

[a] *p*-value is derived from a F-test that measures the quality of curve fit
[b] SD is the standard deviation of sample activities at the lowest tested concentration and values of the DMSO control wells
[c] Class 5 is a special class for samples with activity at zero concentration (zero activity; extrapolated) exceeding 6SD or with zero activity > 3SD and the difference between the maximal change in activity observed in the tested concentration range and zero activity is <3SD

Using compressed efficacy + potency (AUC) and "noise-filtering":

NCATS has used Curvep and weighted AUC

(Hsieh et al. 2015 doi:10.1177/1087057115581317)

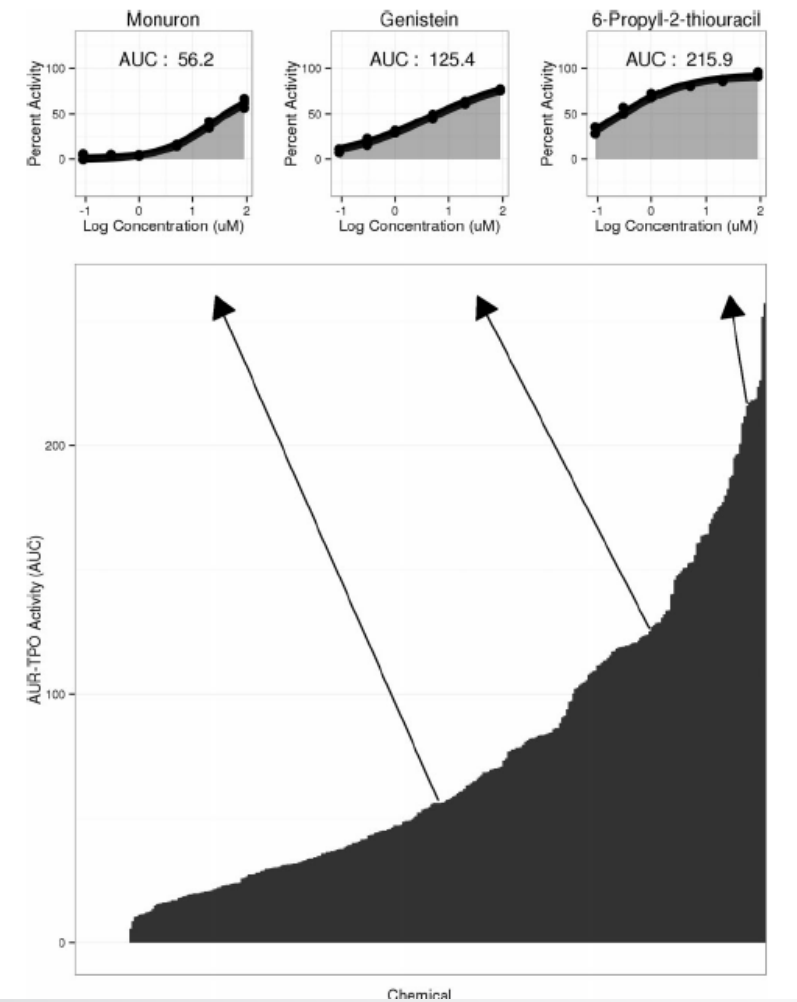


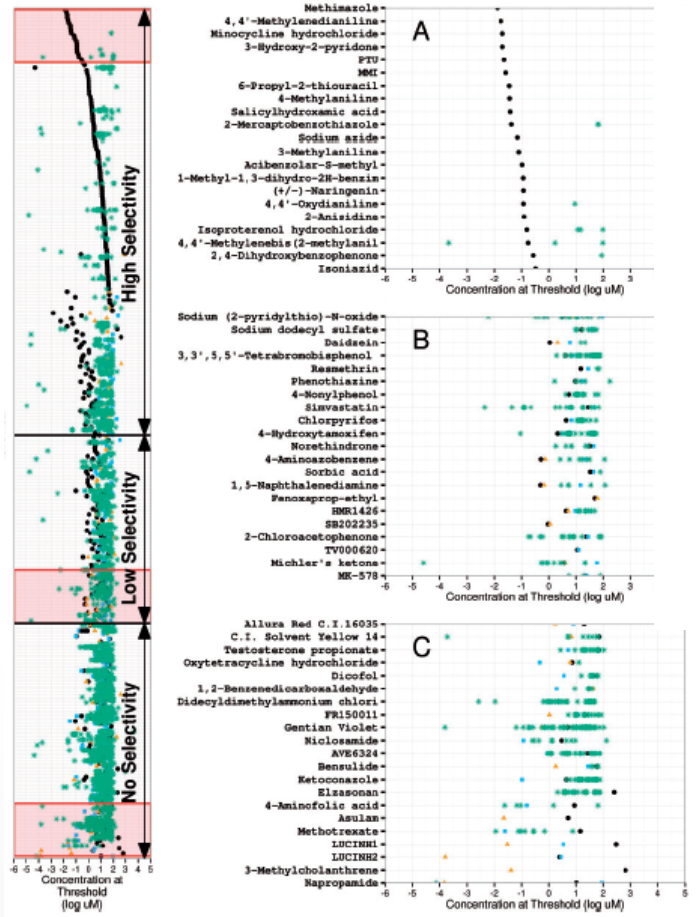
State of the science: ToxCast researchers filter curves, post-release as fit-for-purpose

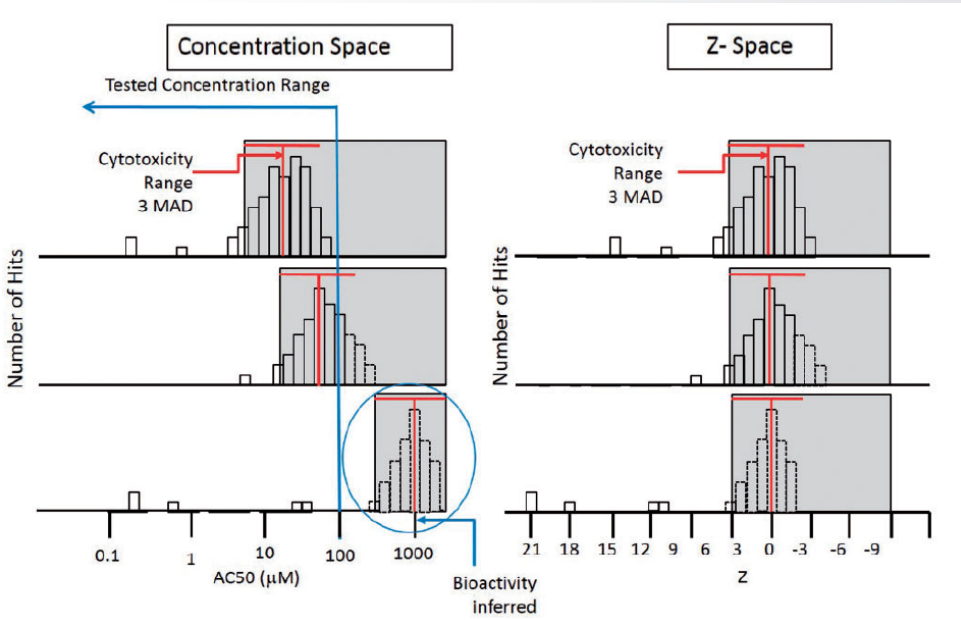
Using AUC and selectivity filtering:

ToxCast research has used AUC and distance from the "burst" or other indicators to indicate selectivity

(Paul-Friedman et al. 2016 doi: 10.1093/toxsci/kfw034, Judson et al. 2016 doi: 10.1093/toxsci/kfw092)







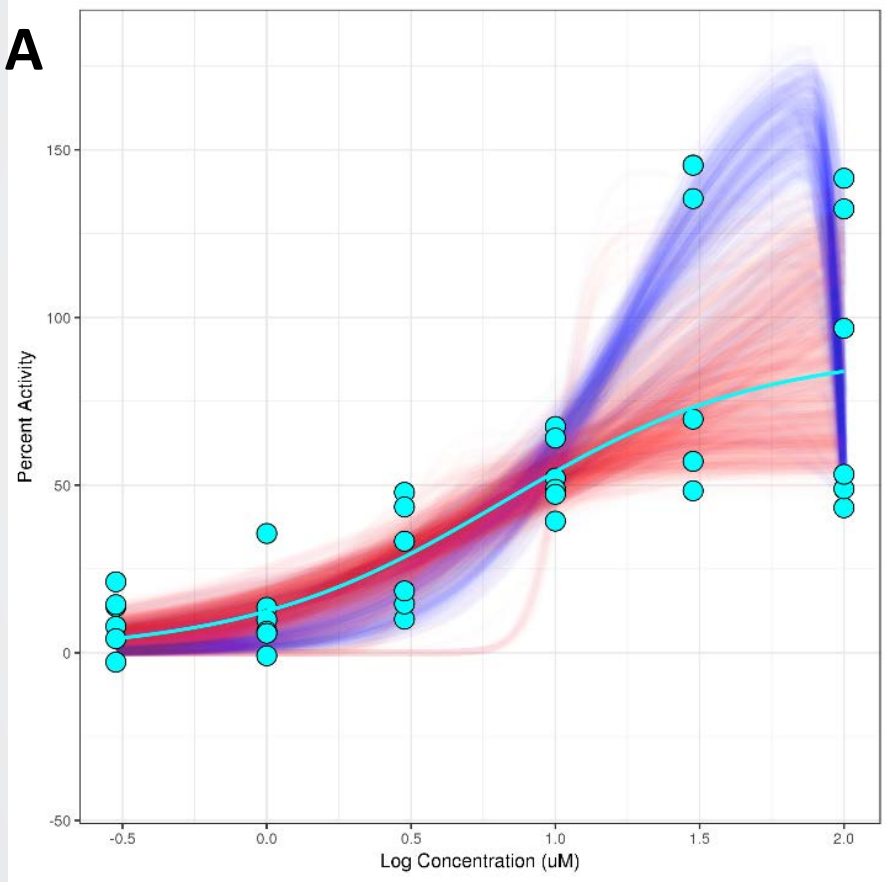
Possible solution: implement toxboot R package (Watt, et al. *in review)* for all of *invitrodb*

- 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.

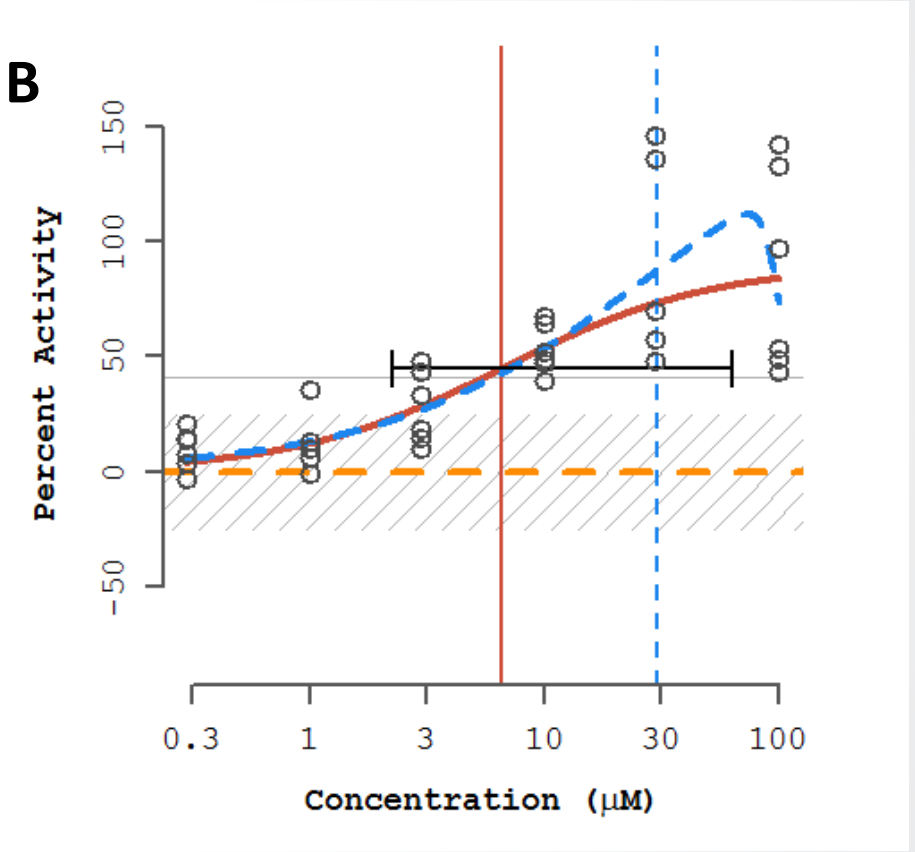
- As hit-calls are binary (positive or negative), they are susceptible to variability and uncertainty in curve-fitting.

- 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.

A bootstrap resampling approach to defining possible curve fits



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).



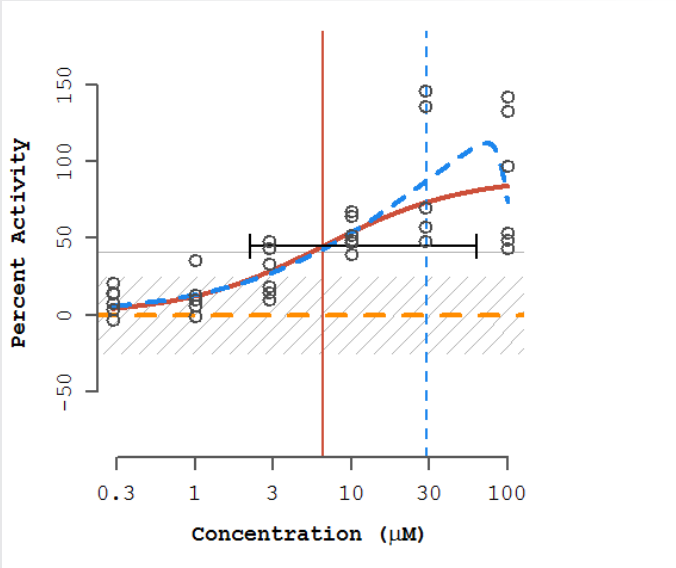
The same plot from Panel A is shown as a tcpl level 7 plot with the added AC50 95% confidence interval width added to summarize the toxboot uncertainty estimation.

Early implementation: Challenges and solutions

- **Challenge 1: Computational time.** With 2.2 million concentration response series in invitrodb_v2, it would take ~10 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 are stored back to a new level 7 table in invitrodb (pre-release).

Preparing for the next release of invitrodb: populating level 7 (mc7)

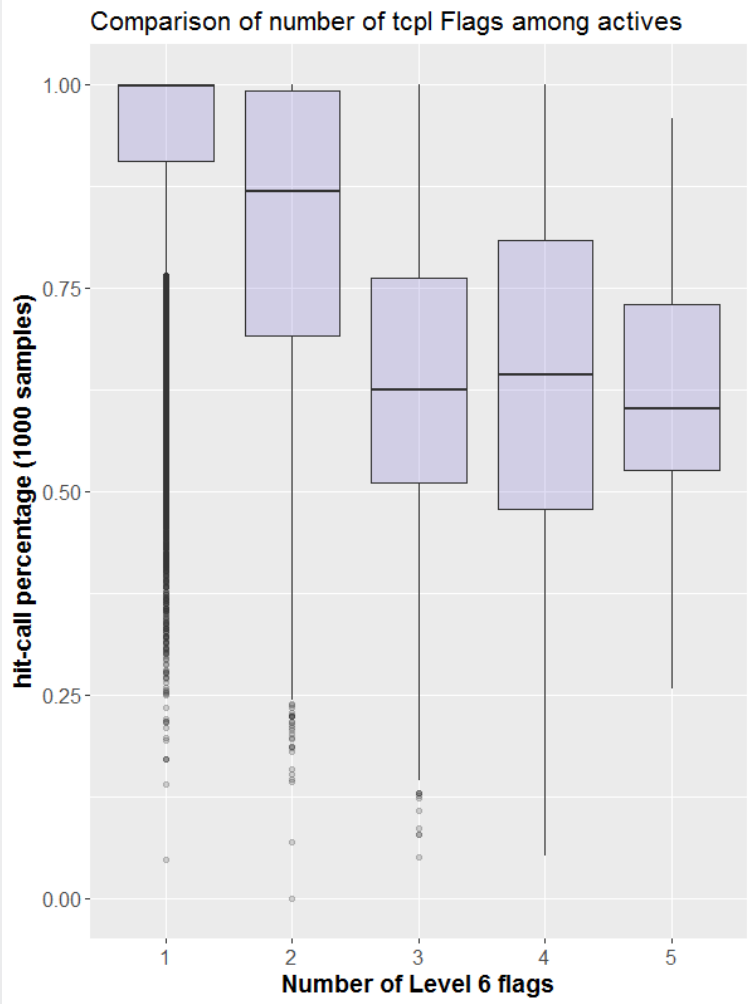
Example illustrations of toxboot results




```
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        348.64
PROB:  0             0.8           0.2
RMSE:  64.8          27.91         26.62

MAX_MEAN: 100        MAX_MED: 103        BMAD: 8.17

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

FLAGS:

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

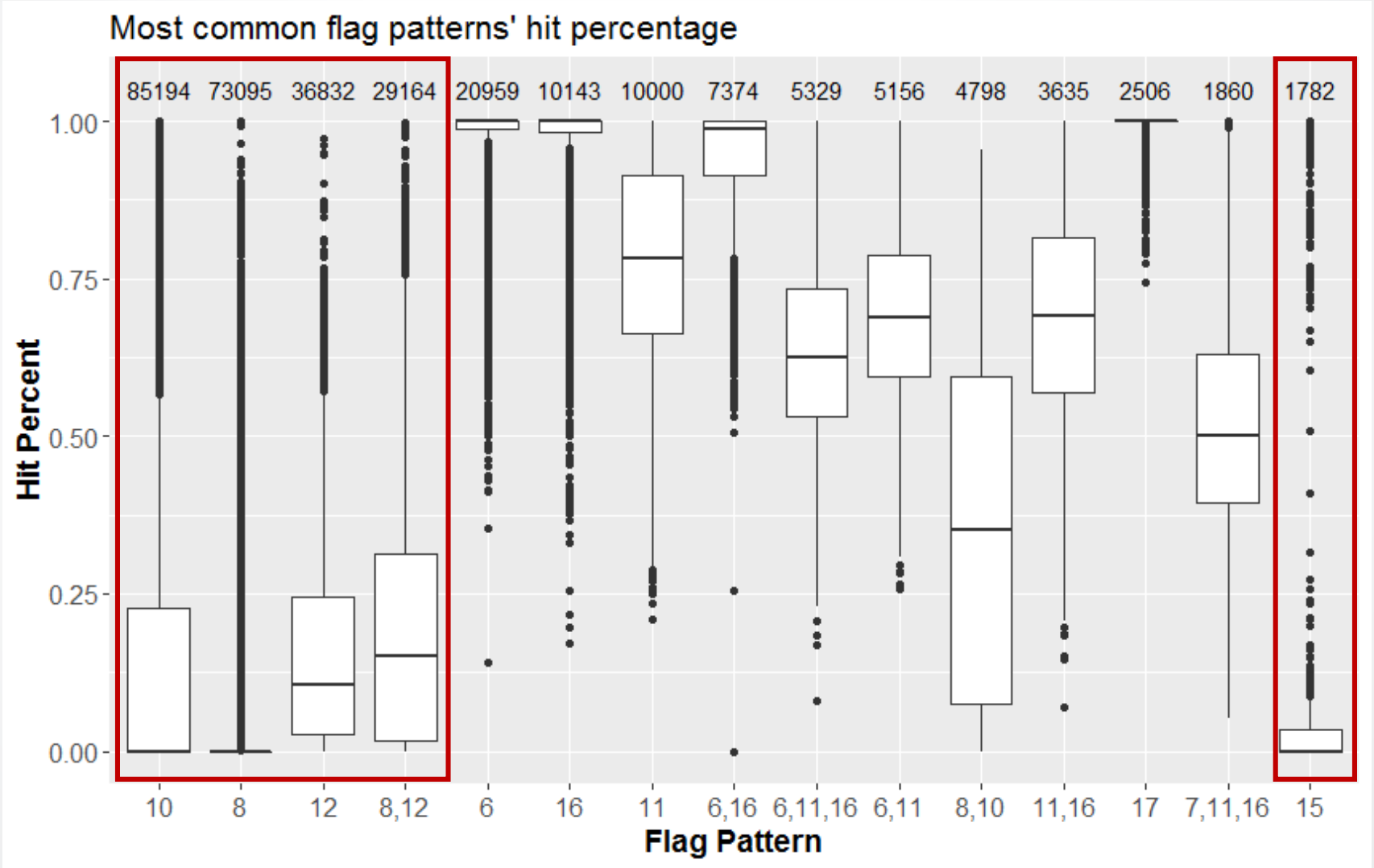
Stored Parameter	Description
Hit_pct	Hit Percentage
Modl_ga_min, Modl_ga_max, Modl_ga_delta	Lower, upper, and width of the AC50 confidence interval
Modl_ga_med	Median AC50 calculated from bootstrapping
Modl_gw_med	Median hill coefficient calculated from bootstrapping

Filtering by caution flags: may work



- Curves with multiple flags have a wide range of hit percents, but the median hit percent for 3+ flags appears to be ~60-65%...
- So filtering by flag sum + hit-percent may remove “worst,” but may not be a complete approach.

Specific flags: some patterns correspond to less reproducible fits than others? Still not “perfect”



These 15 flag patterns cover over 95% of the different types of flag patterns in invitrodb_v2.

Part III: Conclusions

- We are actively quantifying uncertainty in the tcpl-derived curve fits.
- Use of this information may be fit-for-purpose, and so summary information for the user will be stored in mc7.
- Simple rules may work for filtering curve fits (flags, fitc, and hit-percent) depending on the purpose, but it may be ideal to try to build a model using these and other features.
- It may be that combinations of these features are more informative locally (e.g., for one assay or technology), rather than globally across the database.

Acknowledgements

EPA-ORD-NCCT and Tox21 collaborators

Derik Haggard (NCCT)
Jason Brown (NCCT)
Eric D. Watt, now at Pfizer
Woody Setzer (NCCT)
Richard Judson (NCCT)

EPA's National Center for Computational Toxicology