Software and Database Enhancements to ToxCast™ for Accessible Bioactivity Data for Toxicology

Madison Feshuk[1], Jason Brown[1], Sarah Davidson-Fritz[2], Katie Paul Friedman[1]
Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency,
Research Triangle Park, NC[1] Cincinnati, OH[2]

CompTox Communities of Practice January 25 Webinar

Outline & Disclaimer

- ToxCast Overview & Rationale for Updates
- Summary of Updates
- Activity & Potency Estimates
- Version Comparison: How do ToxCast Pipeline (tcpl) updates affect results?
- Exploring ToxCast Data

The views expressed in this presentation are those of the presenter and do not necessarily reflect the views or policies of the US Environmental Protection Agency.

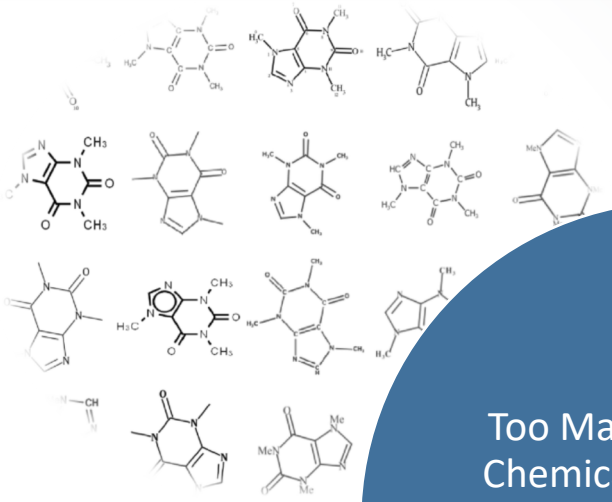
Company or product names do not constitute endorsement by US EPA.

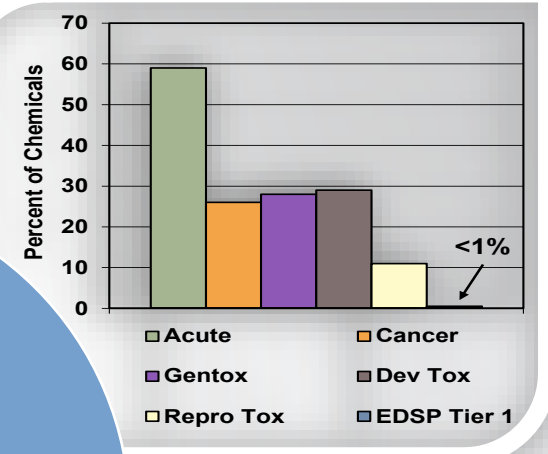
ToxCast Overview

Katie Paul Friedman

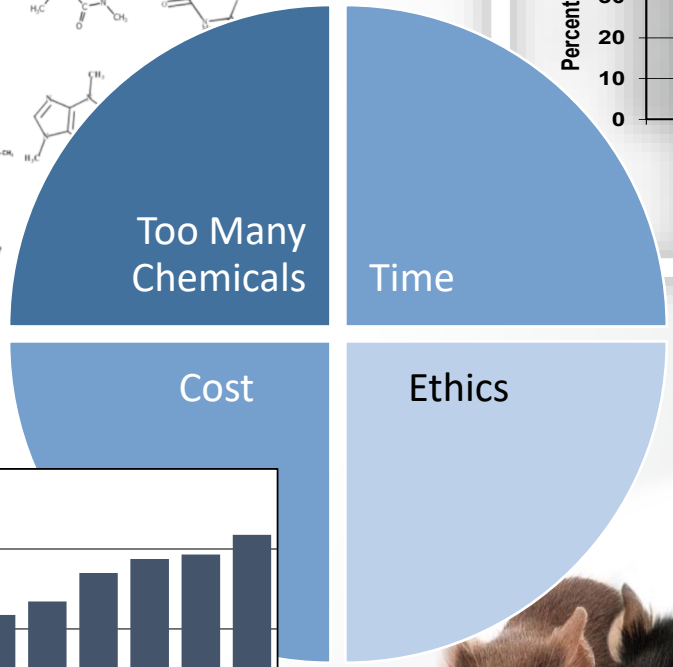
Need for Predictive Toxicology in Next-Gen Risk Assessment

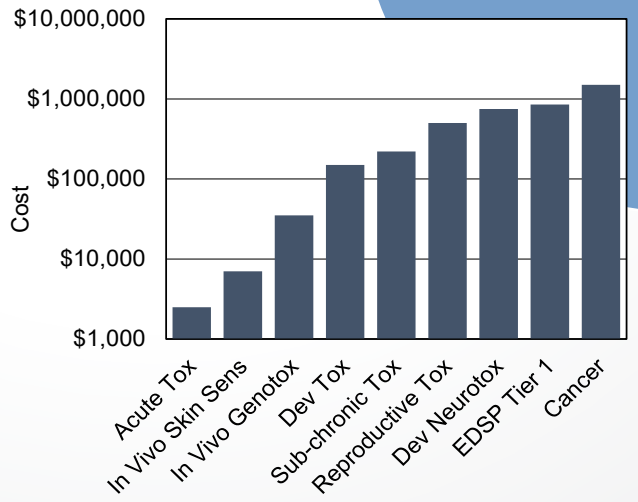
- There are a number of limitations to traditional toxicology testing
- EPA needs rapid and efficient methods to prioritize, evaluate and regulate thousands of chemicals in commerce
- CompTox Blueprint outlines a tiered testing strategy for hazard characterization (Thomas *et al.*, 2019)
  - Tier 1: Broad profiling, high content assays
  - Tier 2: Targeted *in vitro* assays (e.g. ***ToxCast***)
  - Tier 3: Confirmation using assays of greater biological complexity (e.g. ***ToxCast***)





Modified from Judson *et al.*, EHP 2010



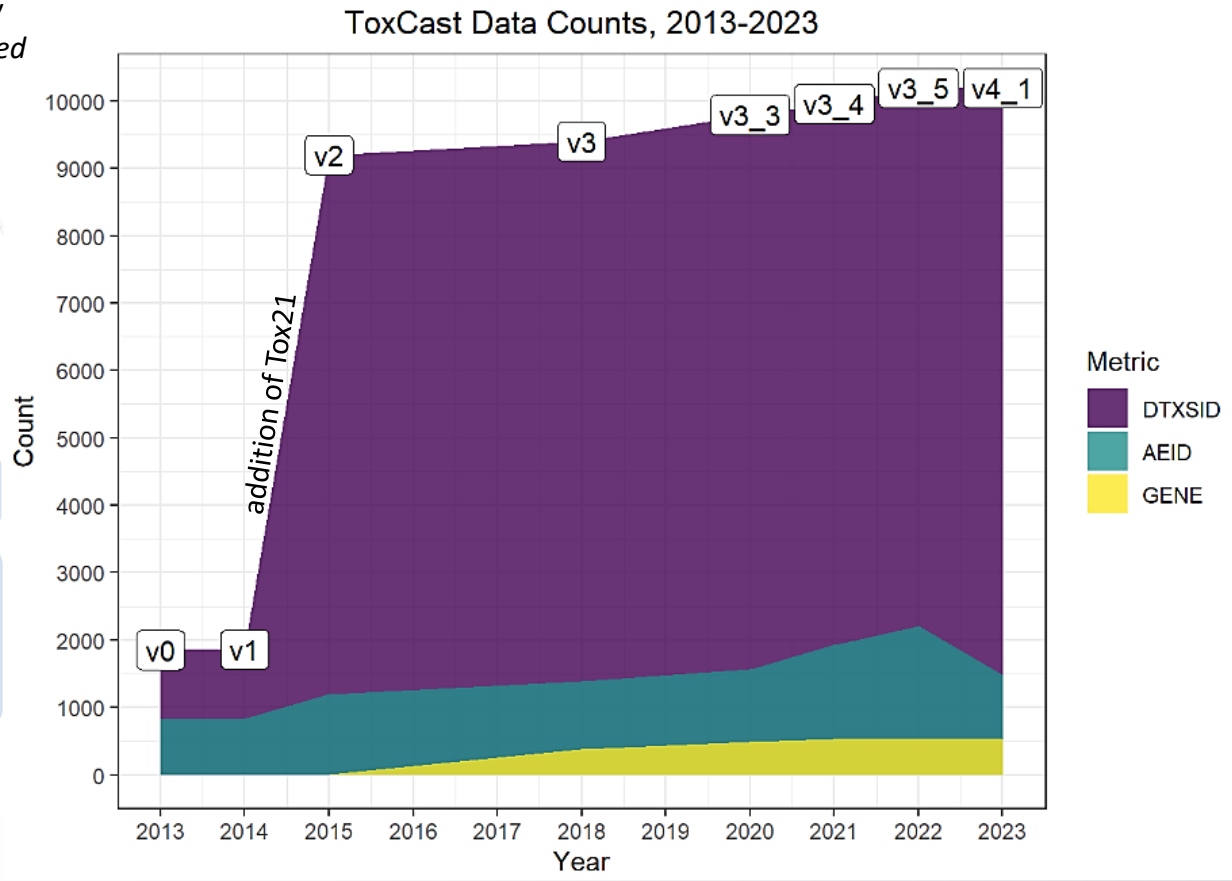


ToxCast Database Coverage

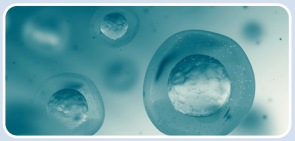
The **Toxicity Forecaster (ToxCast)** program curates and makes publicly available targeted bioactivity screening data. Latest database release (v4.1) includes:

- 26 Assay Sources
- 623 Unique Assays
- 1496 Unique Endpoints
- 9559 Chemicals

Including a TOX21 assay source for data generated by the TOX21 program



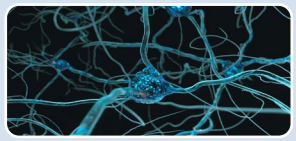
Diverse biology with ***over 500 mapped gene targets***, including:



Endocrine-Related: Estrogen Receptor, Androgen Receptor, Thyroid, Steroidogenesis

Cellular Signaling Pathways: Cytotoxicity, Proliferation, Stress, Mitochondrial Toxicity

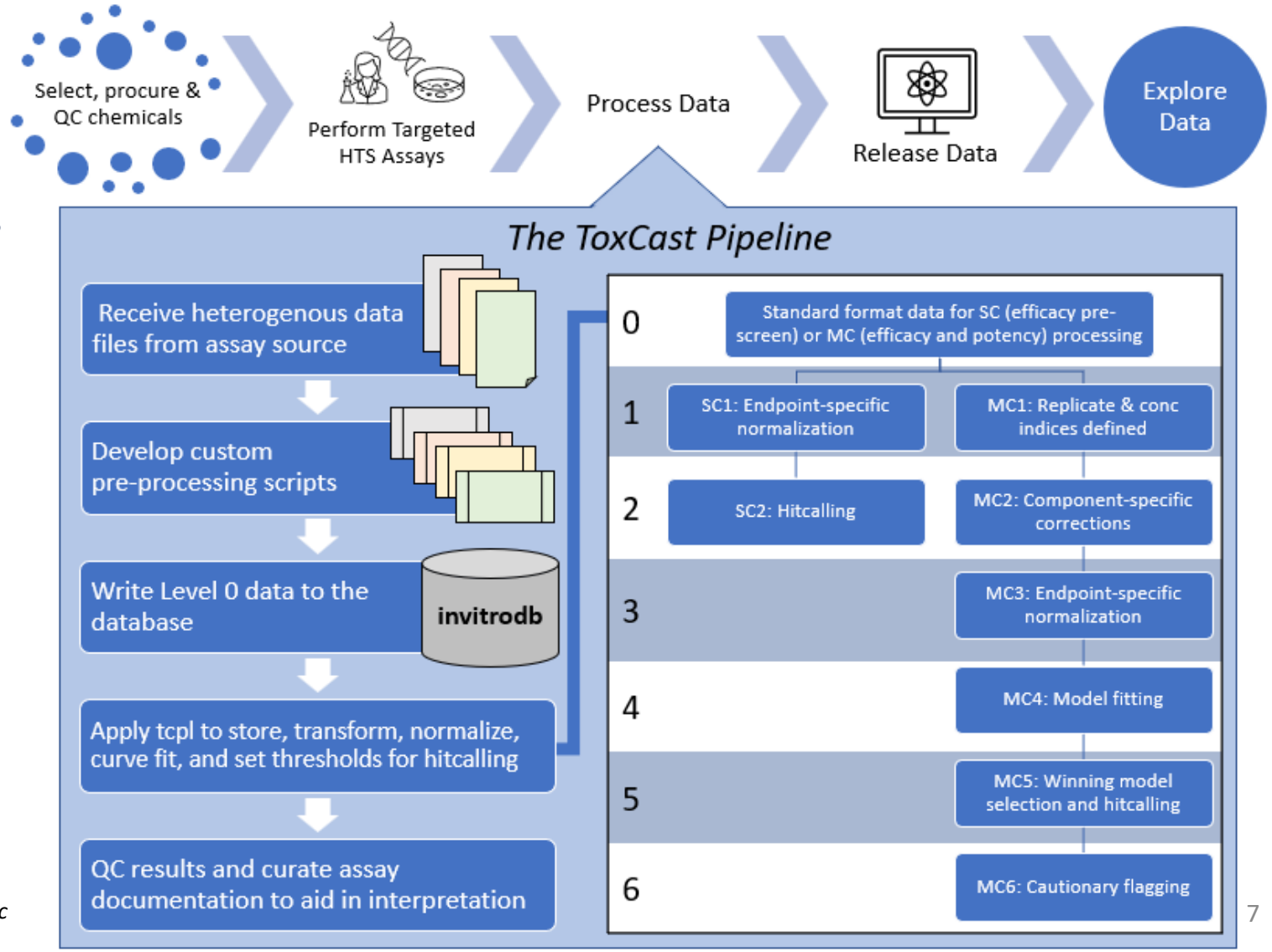
Protein Interactions: Receptors, Transporters, Ion Channels, Enzymes



Complex Responses, e.g. Immune Response, Development, Neurotoxicity, etc.

Process Overview

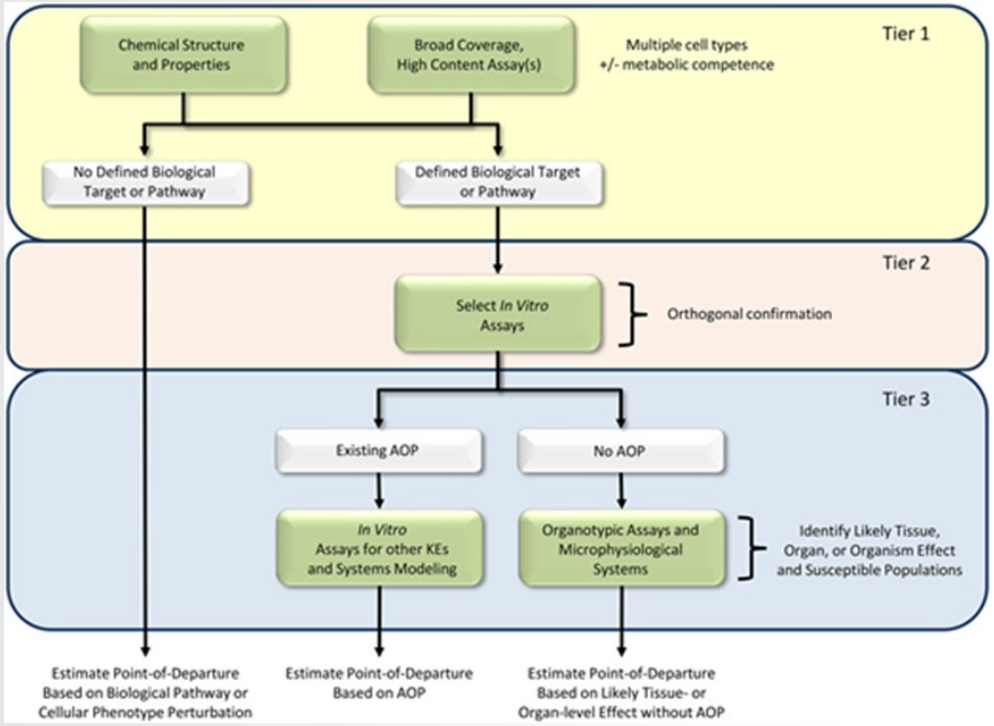
- EPA will **Select, Procure, and QC chemicals,** then ship the chemicals to different *assay sources*, such as EPA labs, contract vendors, or other partners*, who **Perform Targeted HTS Assays**
- Heterogeneous assays = heterogenous assay readouts
- Once the output files are received from the assay source, ToxCast team can **Process Data** with tcpl
  - The ToxCast Pipeline (tcpl) R software package to populate its linked MySQL database, invitrodb
  - Tcpl is a flexible analysis pipeline capable of processing and storing large volumes of data in addition to all processing decisions and metadata
- After additional QC and curation, the ToxCast team **Release data** annually via the ToxCast Downloadable Data page
- EPA & the public can **Explore data** through data downloads or via the CompTox Chemicals Dashboard



**initiated via Material Transfer Agreements (MTAs) according to the Agency's strategic research needs under Chemical Safety for Sustainability (CSS) Research Program*

Rationale to adding tcplfit2 dependency

How to connect between tiers when the curve-fitting may be different between them?



Thomas *et al.* 2019

- Use a single curve-fitting approach across tiers, encoded by *tcplfit2*, to enable comparison across all bioactivity data
- Models in *tcplfit2* are based on models in BMDExpress2
- Flexibility to add more curve-fitting models in the future to better capture the varied response behavior observed in *in vitro* NAMs
- Reduction in data redundancy (no more “up” and “dn” refitting) to simplify interpretation, annotations, and modeling tasks that utilize ToxCast data as input
- Improve interoperability of all bioactivity data

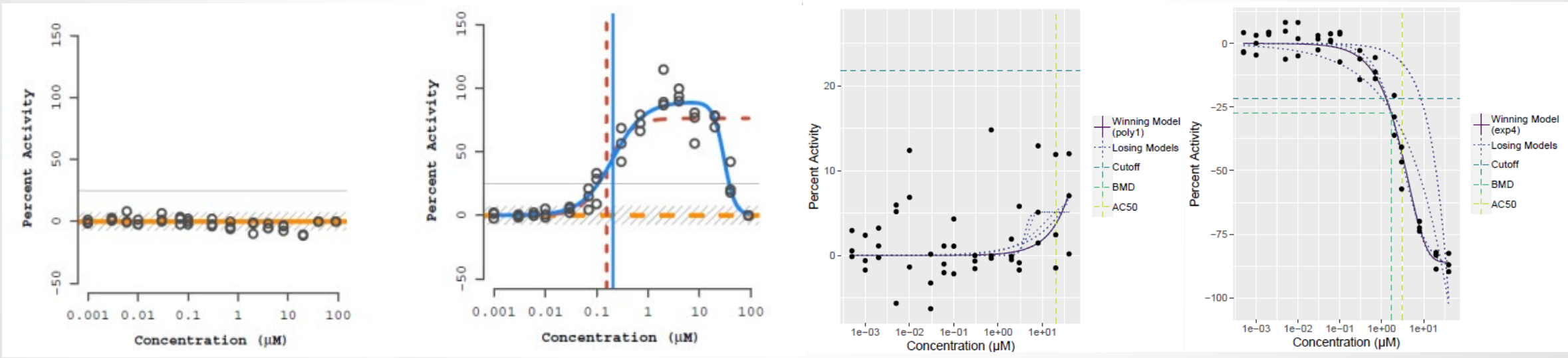
Together, these updates to tcpl and invitrodb improve the utility of ToxCast data within an integrated NAM strategy and unified open-source software approach.

Updates

Jason Brown

Updates from tcpl v2.0 to tcpl v3.0

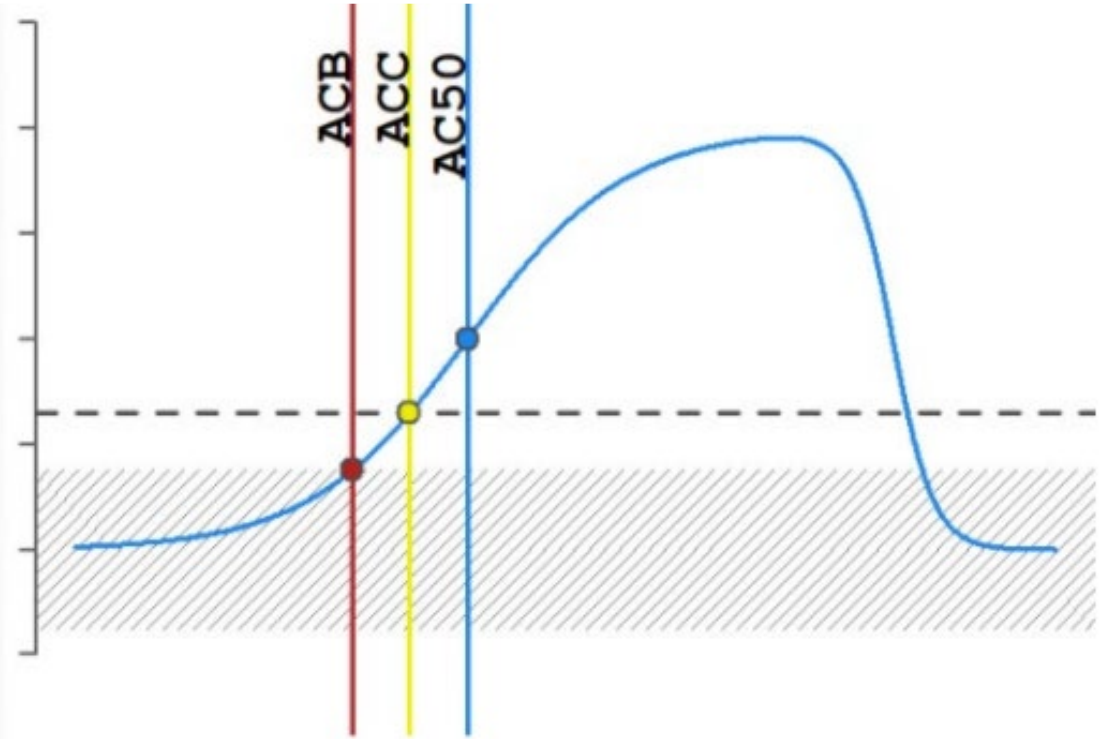
Enhancement	InvitroDB v3.5 and *Tcpl* v2.0	InvitroDB v4.0 and *Tcpl* v3.0
Curve-fitting models	Models included constant, Hill, and gain-loss.	In addition to constant, Hill, and gain-loss, models included Polynomial 1 (Linear), Polynomial 2 (Quadratic), Power, Exponential 2, Exponential 3, Exponential 4, and Exponential 5 based on BMDExpress and encoded by R package dependency *tcplfit2*
Plotting	Several functions were used to produce the different plotting outputs.	tcplPlot() allows for interactive, yet consistent visualization of concentration-response curves.
Activity hit calls	Hit call was discrete: 0 = negative, 1 = positive, -1 = Unable to fit (usually due to fewer than 4 concentrations).	Hit call is continuous as the product of three proportional weights: median response and top of model both exceed the cutoff, and the winning model is not fit to background noise (Akaike Information Criterion of winning model is less than that of the constant model).

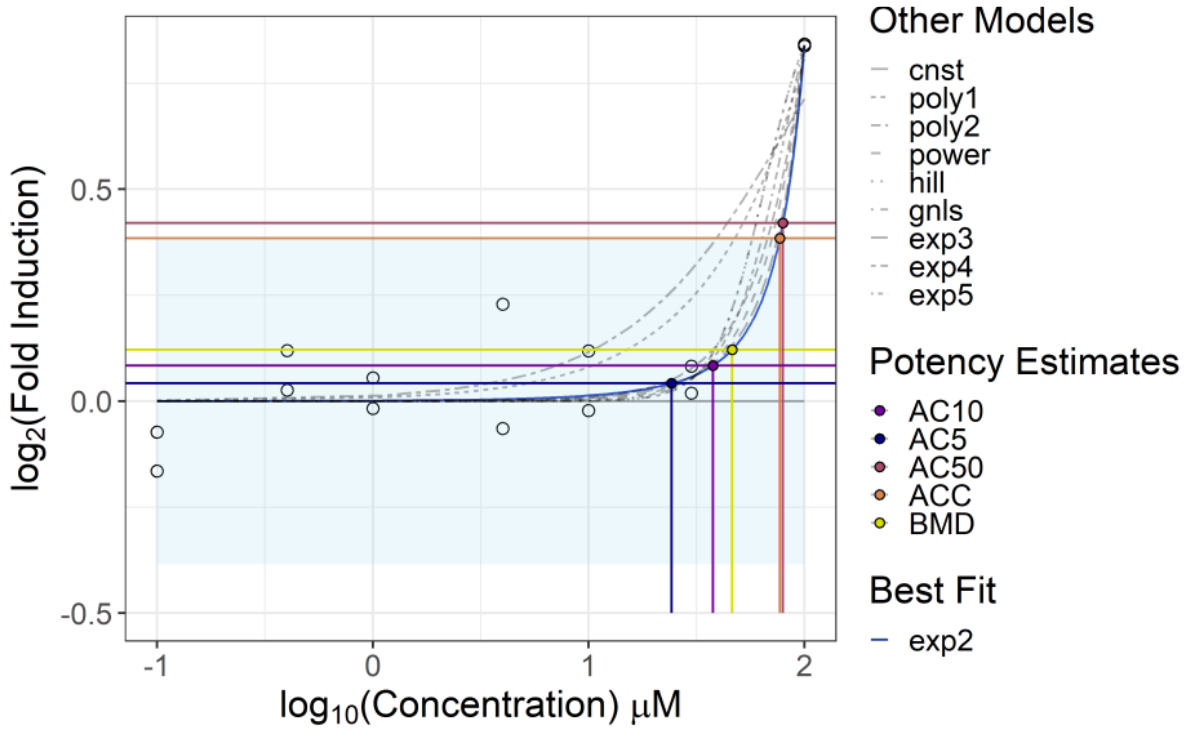


Hit call **0** **1** **9.6e-09** **1**

Updates from tcpl v2.0 to tcpl v3.0

Enhancement	InvitroDB v3.5 and *Tcpl* v2.0	InvitroDB v4.0 and *Tcpl* v3.0
Potency estimates	Point of departure potency estimates were based on modelled active concentration series, including ACB (activity concentration at baseline, 3bmad), ACC (activity concentration at cutoff), and AC50 (activity concentration at 50% of maximal response)	Based on models within the program BMDExpress, tcplfit2 modelling outputs new potency and uncertainty estimates related to a benchmark dose (BMD) as defined by the Benchmark Response (BMR) level in addition to the ACC, AC50, etc.





Updates from tcpl v2.0 to tcpl v3.0

Enhancement	InvitroDB v3.5 and *Tcpl* v2.0	InvitroDB v4.0+ and *Tcpl* v3.0+
Stand-alone pipelining	In addition to connecting to a *tcpl* database, *tcplLite* connection would create flat files structured like invitroDB for stand-alone pipelining applications	tcplLite is no longer supported by *tcpl. tcplFit2,* however, can be used for stand-alone applications, available at https://cran.r-project.org/package=tcplfit2.
Endpoint structure and annotation	*Tcpl* only fit in the positive analysis direction therefore dual endpoints were registered to capture gain and loss of signal.	Given bidirectional fitting, a single endpoint is sufficient to capture both gain and loss of signal. Many endpoints were removed and/or renamed, and annotations were updated to reflect this paradigm shift. Continued curation efforts enable better data aggregation.
Schema changes	Processed data was previously stored in "wide" format with a fixed number of columns in the Level 4 (mc4) and Level 5 (mc5) tables based on three curve-fitting models.	Complete tcplFit2 model parameters are captured within the mc4_param and mc5_param tables, allowing for generic fitting and hit calling, with summary-level statistics now only stored in mc4 and mc5.
Fit Categories	Fit categories (fitc) were based on *the winning model*, active or inactive designation based on hitc, efficacy, and relationship between the AC50 and the concentration range screened.	A more generic approach to fitc enables the addition of any future curve-fitting models, where fitc is largely based upon the relative efficacy and, in the case of actives, the location of the AC50 and concentration at 95% activity compared to the tested concentration range.
Cautionary Flags	Flags were programmatically generated to indicate characteristics of a curve that need extra attention or potential anomalies in the curve or data.	Many of the flags from the past versions of tcpl are re-implemented, with updates to the coded logic largely to address the introduction of the BMD, bidirectional fitting, and the continuous hitc.

Activity & Potency Estimates

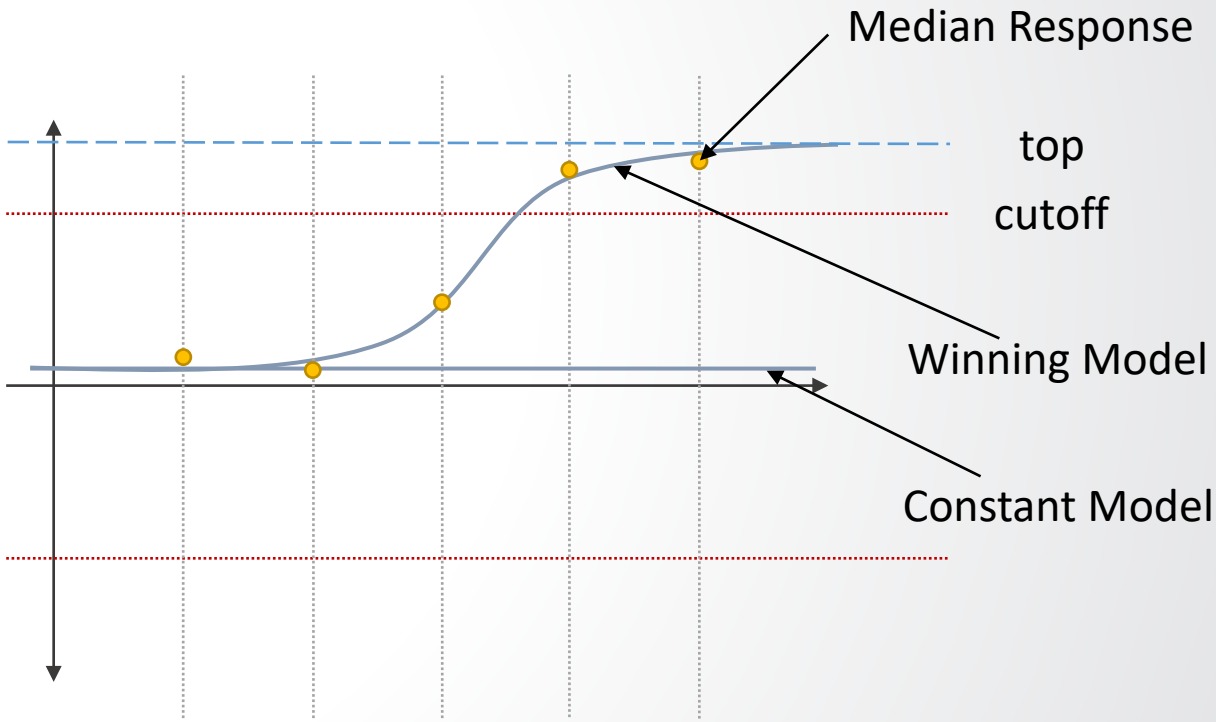
Sarah Davidson-Fritz

Continuous Activity Hit Call (Hitc)

Activity of concentration-response curves in tcpl v3.1 are indicated by the estimated continuous hitc, which is the product of the three proportional weights:

- p_1: "the winning AIC value is less than that of the constant model"
  - Determine whether the constant model – if allowed to win – is a better fit than the winning model – i.e. is the winning model essentially flat or not.
- p_2: "at least one median response is greater than the cutoff"
  - At least one dose group has a central tendency of the response values "outside" the cutoff band (consider bi-directional).
  - Response is greater than cutoff in "+" direction and less than cutoff in "–" direction.
- p_3: "the top of the fitted curve is above the cutoff"
  - Determine whether the predicted maximal response exceeds the cutoff, i.e. the response corresponding to the effect size of interest.

Continuous hit call estimates are between 0 and 1, where values > 0.9 indicate "active" responses.



**tcpl v3+ depends on R package tcplfit2 for curve-fitting and hitcall logic*

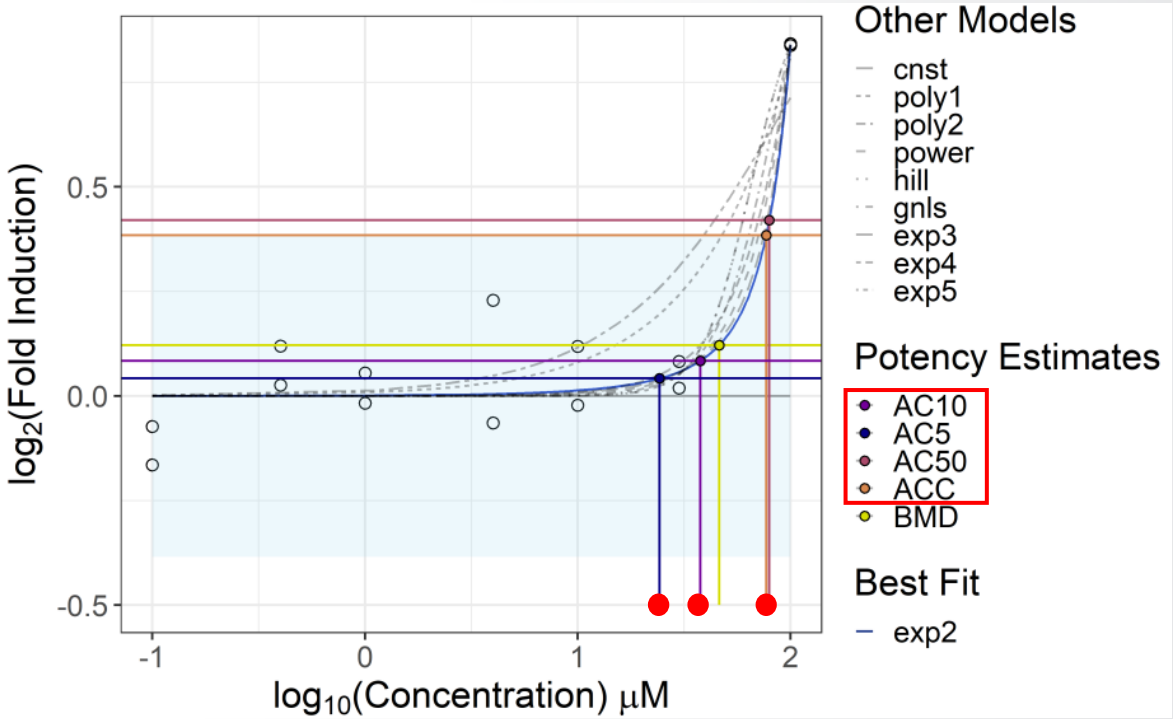
Activity Concentrations

An activity concentration is the estimated concentration inducing a specified level of response (activity).

tcpl v3.1 estimates and tracks several different activity concentrations.

Activity Concentration (uM)	Specified Level of Response
AC5	5% of the maximal response
AC10	10% of the maximal response
AC50	50% of the maximal response
ACC	Response at the user-defined cutoff (threshold)

Additional potency metrics (not shown) are also computed and stored at level 5.



Benchmark Dose (BMD)

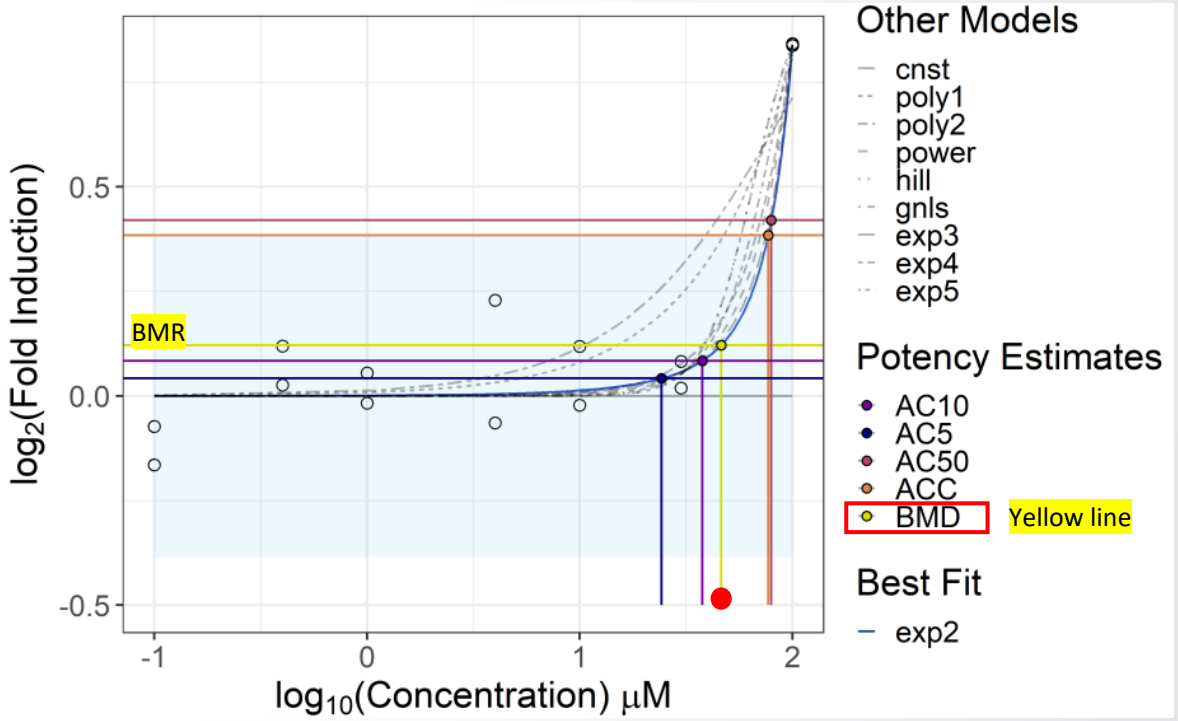
The benchmark dose (BMD) is the concentration inducing a specified benchmark response (BMR).

tcpl v3.1 uses the following definitions and assumptions for setting the BMR:

- BMR is a change from the mean response at baseline ($\mu(b)$) by some multiple (c) of the standard deviation of the baseline ($sd(b)$).

$$\mu(b) + c * sd(b) = BMR = \mu(BMD)$$

- Here, the baseline (b) is defined as samples from the two lowest concentrations across chemicals within an assay endpoint and the $c = 1.349$[a].



[a]Yang, L., Allen, B.C. & Thomas, R.S. BMDExpress: a software tool for the benchmark dose analyses of genomic data. *BMC Genomics* **8**, 387 (2007). https://doi.org/10.1186/1471-2164-8-387

Version Comparison

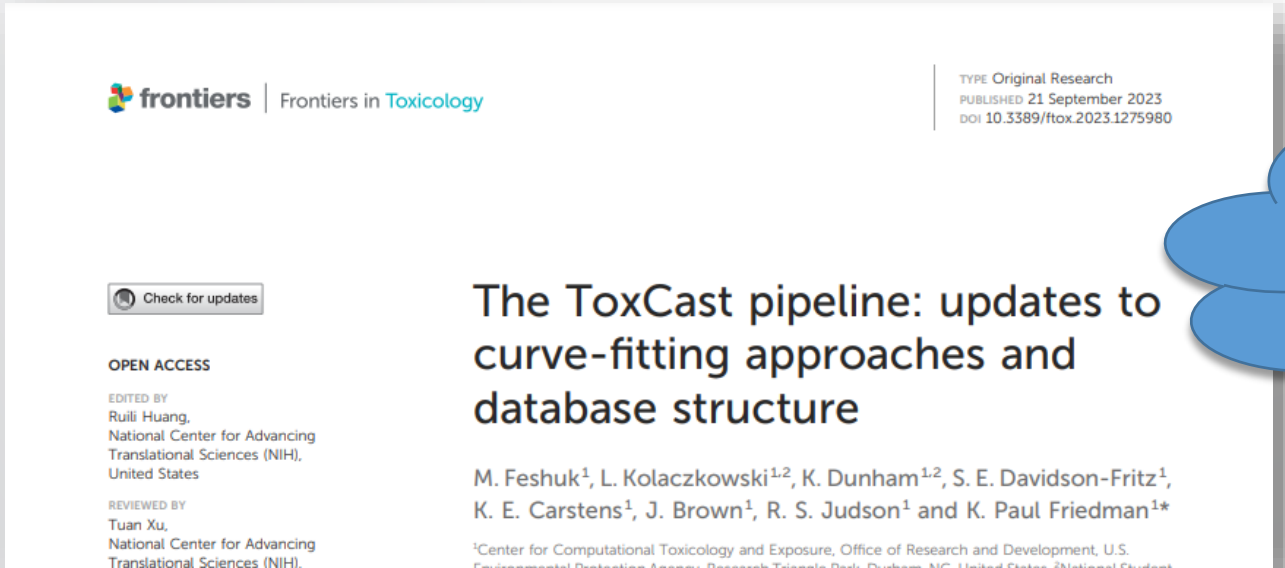
Madison Feshuk

Full publication available here: Feshuk, M., Kolaczkowski, L., Dunham, K., Davidson-Fritz, S. E., Carstens, K. E., Brown, J., Judson, R. S., & Paul Friedman, K. (2023). The ToxCast pipeline: updates to curve-fitting approaches and database structure. *Frontiers in toxicology*, *5*, 1275980. https://doi.org/10.3389/ftox.2023.1275980

Overview

To understand the impacts of tcpl updates on ToxCast data, we compared:

- **invitrodb v3.5**, processed using tcpl v2.1.0, and
- **invitrodb v4.0**, which includes the same data as invitrodb v3.5 but reprocessed with tcpl v3.0.1 into an updated database schema to accommodate enhancements



frontiers | Frontiers in Toxicology

TYPE Original Research
PUBLISHED 21 September 2023
DOI 10.3389/ftox.2023.1275980

Check for updates

OPEN ACCESS

EDITED BY
Ruili Huang,
National Center for Advancing Translational Sciences (NIH), United States

REVIEWED BY
Tuan Xu,
National Center for Advancing Translational Sciences (NIH),

The ToxCast pipeline: updates to curve-fitting approaches and database structure

M. Feshuk[1], L. Kolaczkowski[1,2], K. Dunham[1,2], S. E. Davidson-Fritz[1], K. E. Carstens[1], J. Brown[1], R. S. Judson[1] and K. Paul Friedman[1]*

[1]Center for Computational Toxicology and Exposure, Office of Research and Development, U.S.

invitrodb v4.1 is the first public release in the new schema.
invitrodb v4.0 was a beta release to assess impact of updates.

Bidirectional fitting decreased redundancy in endpoints

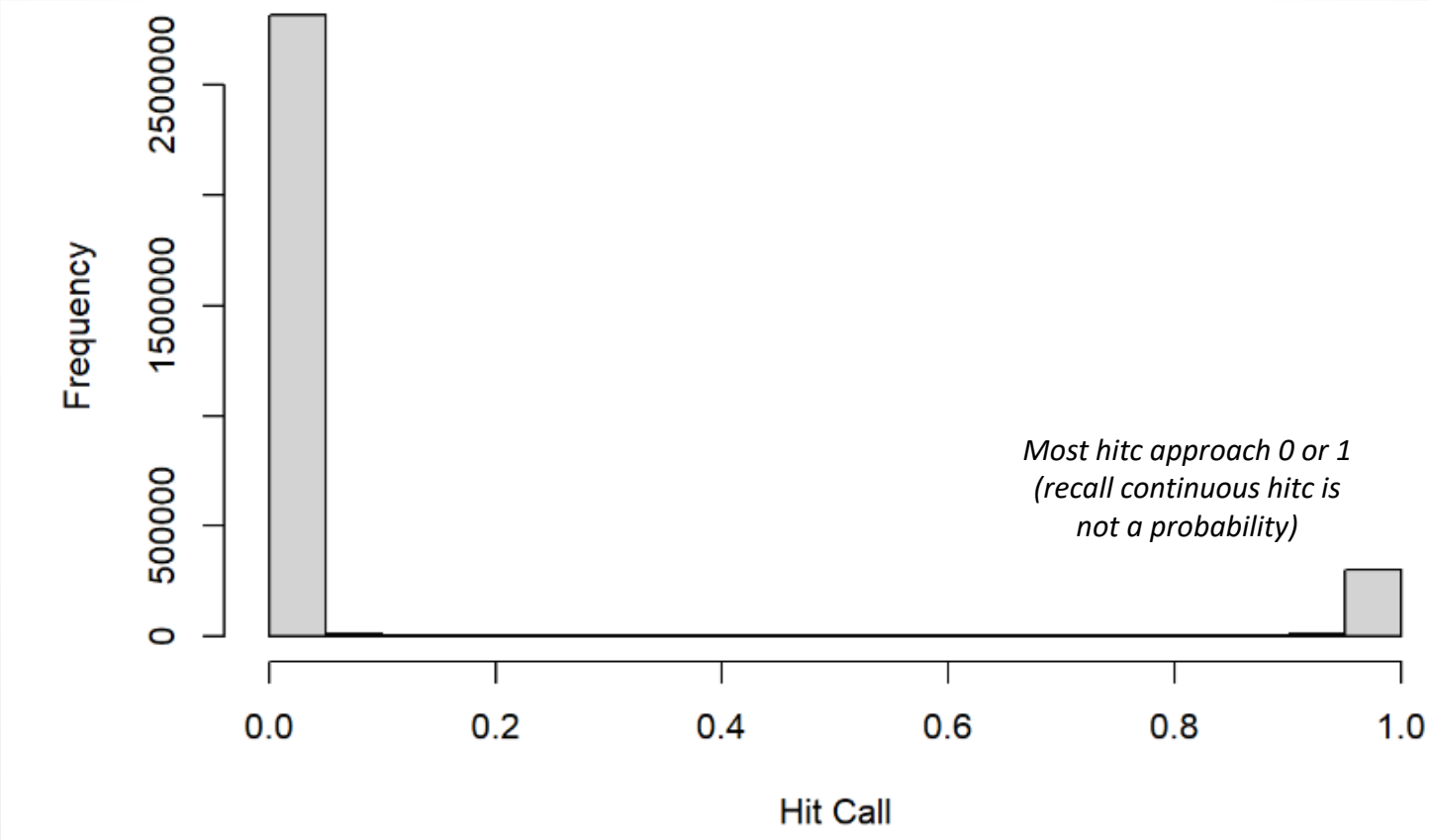
Assay Element	invitrodb v3.5	invitrodb v4.0	Change
Assay Sources	26	26	0
Assays	623	625	2
Assay Components	1499	1496	-3
Assay Component Endpoints	2243	1496	**-747**
Samples: Distinct quantity of chemical procured and screened	46712	46712	0
Chemicals: Unique chemical compounds screened	9541	9541	0
Endpoint-Samples: Combination of unique samples screened per endpoint	3979274	3215442	**-763832**

- Invitrodb v4.0 saw a reduction by 747 endpoints (and 763,832 redundant curves) given bidirectional fitting

Activity hit calls are now continuous

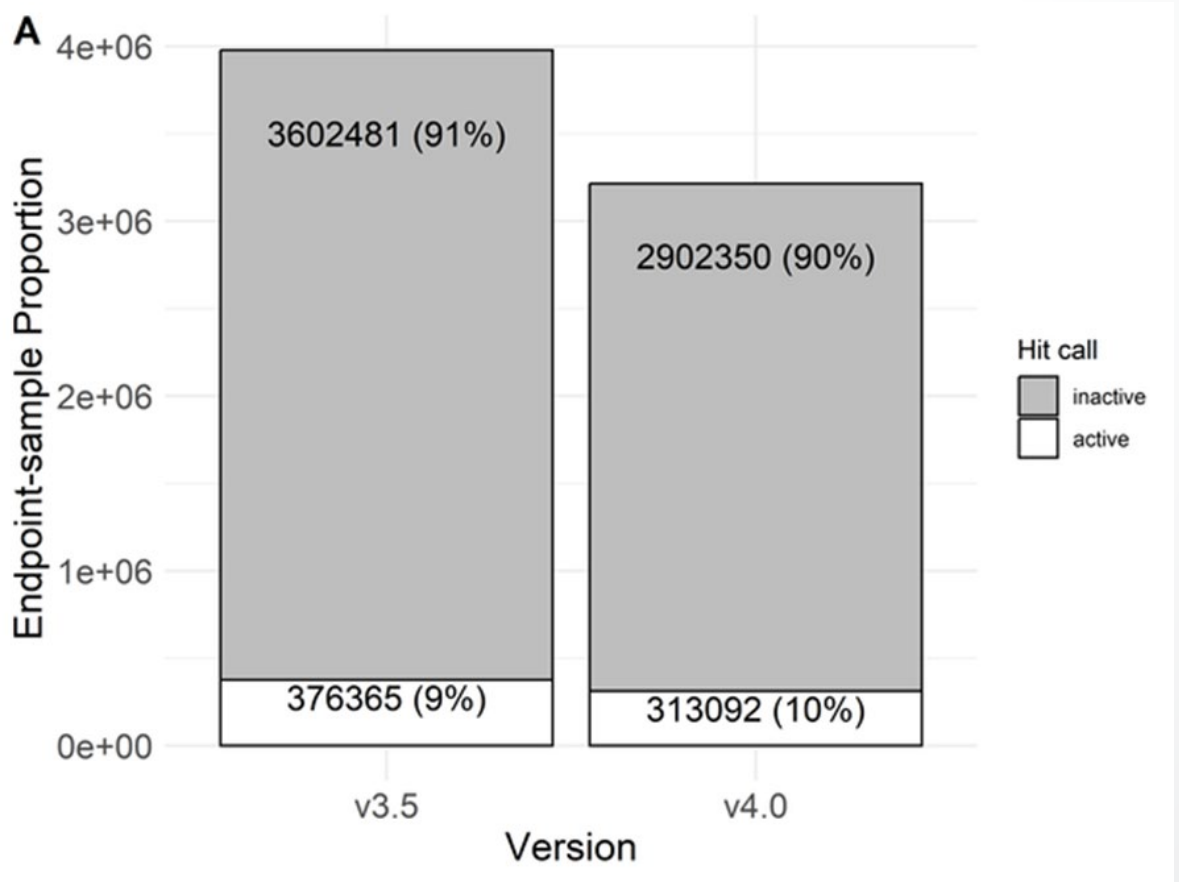
- In invitrodb v3.5, hit call (hitc) were discrete (0,1) whereas in invitrodb v4.0, the hitc is continuous (0-1) product of proportional weights
- For this analysis, we used a threshold for actives: hitc >= 0.90 is active, whereas hitc < 0.90 is inactive

Distribution of hitc in invitrodb v4.0



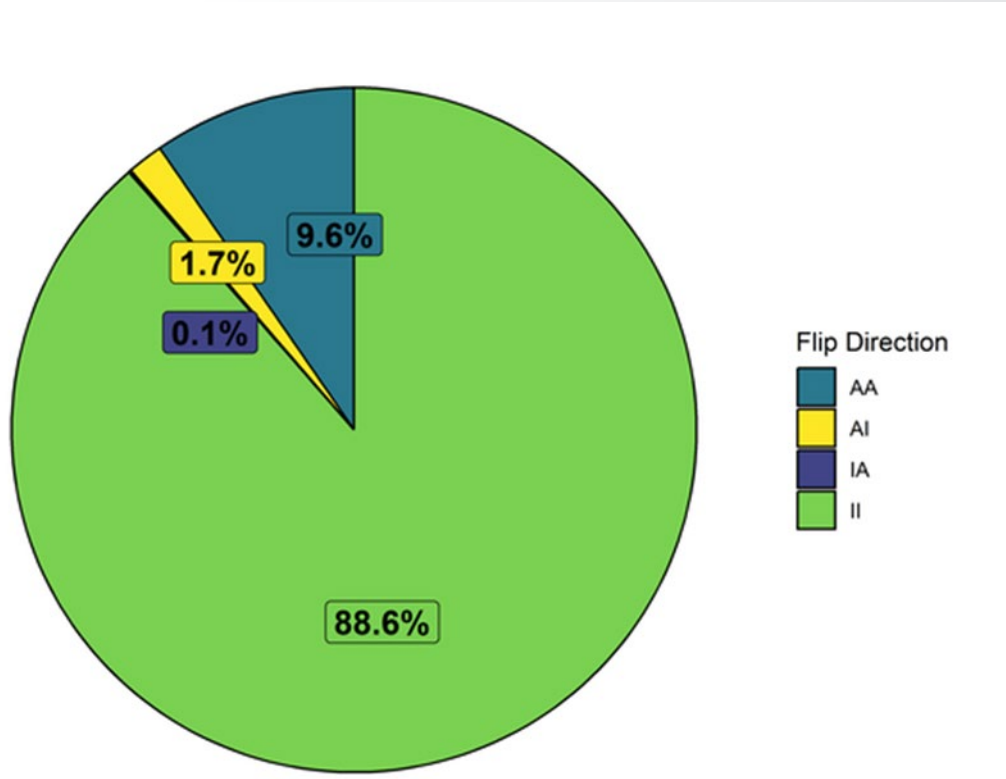
Activity Hit Calls: Proportion

- invitrodb v3.5 included 91% inactive and 9% active hitc whereas invitrodb v4.0 included 90% inactive and 10% active



Activity Hit Calls: Flipped

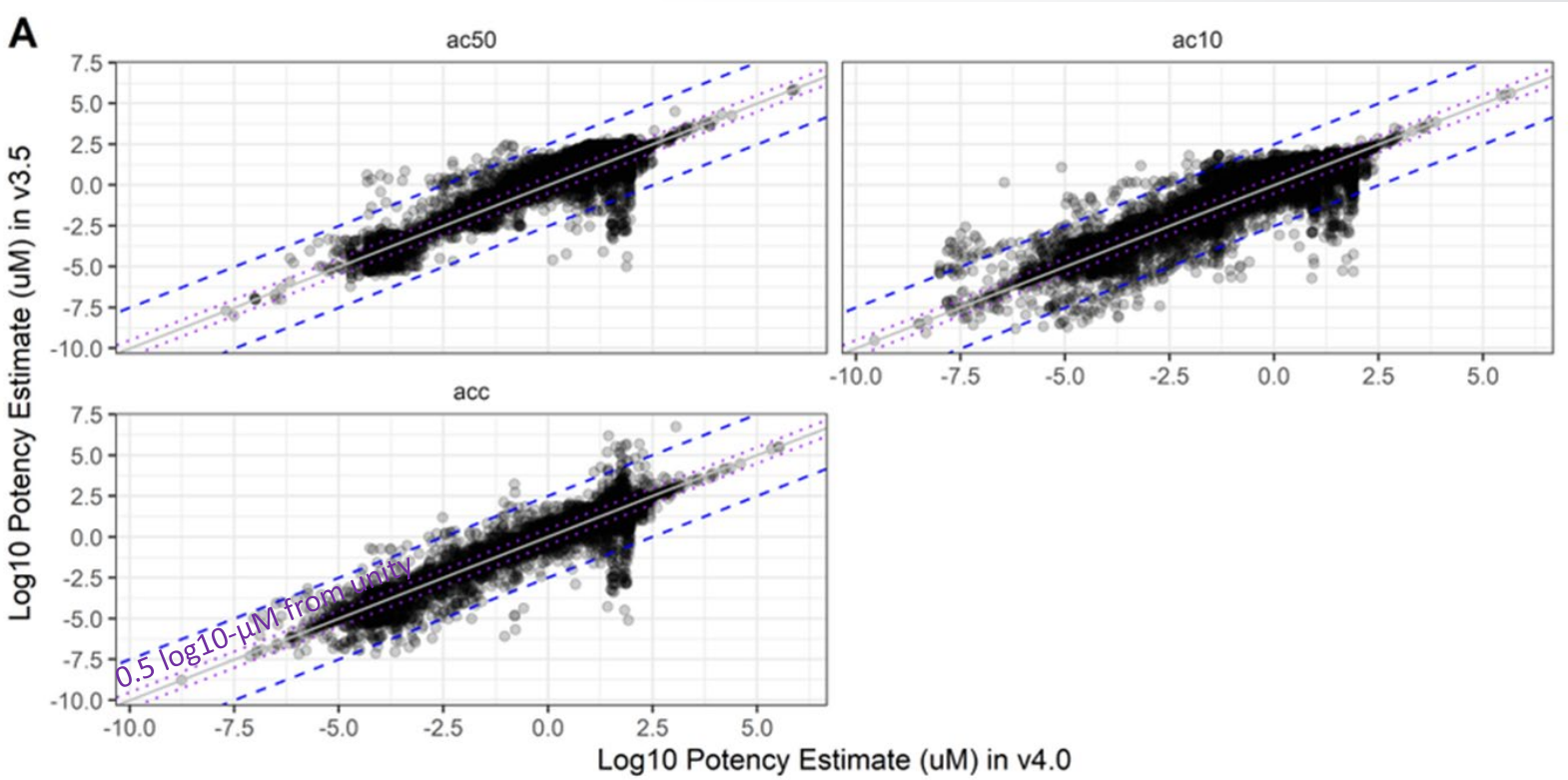
- Potential changes in individual hitc were also evaluated in aggregate by endpoint
- Possible hitc flip directions:
  - AA: active in both invitrodb v3.5 and v4.0
  - AI: active in invitrodb v3.5, but inactive in v4.0
  - II: inactive in both invitrodb v3.5 and v4.0
  - IA: inactive in invitrodb v3.5, but active in v4.0
- 98.2% unchanged (II, AA)
  - In terms of flipped hitc, 1.7% of endpoint-samples were AI and only 0.1% converted to IA



Flipped hitc seemed related to responses with lower efficacy (borderline activity) or activity from a single point/possible noise

Potency

- All potency values (ACC, AC10, and AC50) from invitrodb v3.5 and v4.0 fall within −5 and 2.5 on the log10-µM scale
- ACC and AC50 comparisons largely fall on or within 0.5 log10-µM of the unity line
- RMSD was computed along with bootstrap-resampled 95% confidence interval around these RMSD values, which suggest that **AC10, ACC, and AC50 values were on average 0.28, 0.16, and 0.20 log10-µM different, respectively, between invitrodb versions**



Potency Comparison	Calculation	2.5% Lower Bound	RMSD	97.5% Upper Bound
AC10	v4.0 - v3.5	0.271	0.275	0.279
AC50	v4.0 - v3.5	0.192	0.196	0.200
ACC	v4.0 - v3.5	0.154	0.158	0.163

But wait, there's more!

See full publication for additional version comparison, including:

- Assay source-specific analyses
- Further inspection of flipped hit calls using v3.5 flags, fit categories, and ratio of top over cutoff (i.e., reasons why hit calls may have flipped)
- Changes in the winning models
- Cytotoxicity burst threshold shifts

Full publication available here: Feshuk, M., Kolaczkowski, L., Dunham, K., Davidson-Fritz, S. E., Carstens, K. E., Brown, J., Judson, R. S., & Paul Friedman, K. (2023). The ToxCast pipeline: updates to curve-fitting approaches and database structure. *Frontiers in toxicology*, *5*, 1275980. https://doi.org/10.3389/ftox.2023.1275980

Exploring ToxCast Data

Madison Feshuk

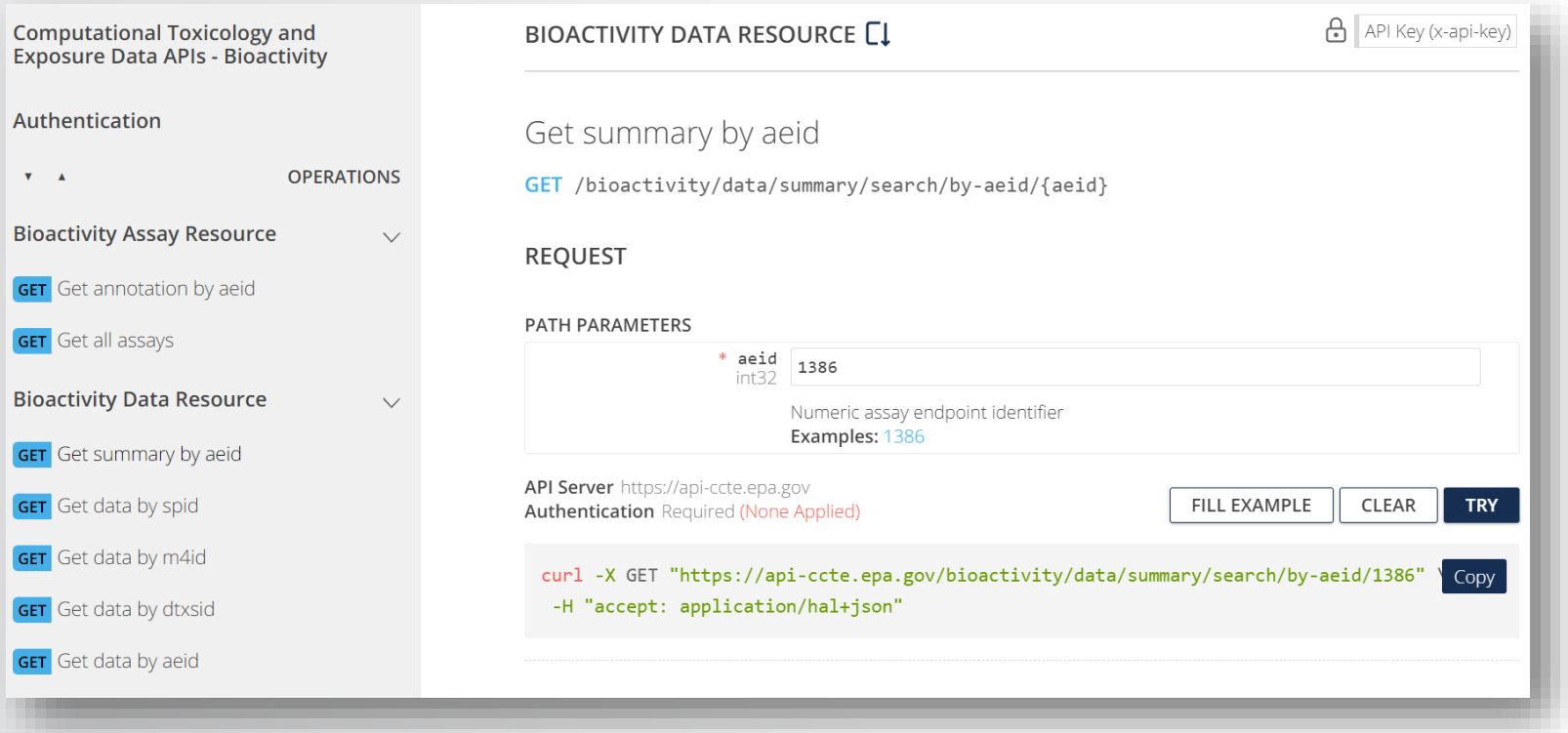
CompTox Chemicals Dashboard(CCD)

https://comptox.epa.gov/dashboard

- CCD's ToxCast bioactivity module presents a view of potency and relative efficacy metrics across ToxCast endpoints for chemicals of interest
- Users can easily sort, filter, and export ToxCast results and assay descriptions
- Notable updates in the CCD v2.3 release (December 2023) include:
  - Data was refreshed to invitrodb v4.1
  - ToxCast Summary tab is now a single tab that combines the previous ToxCast Summary and ToxCast Conc. Response tabs
  - Bioactivity Summary Grid includes v4.1 information in new columns, including benchmark dose (BMD), benchmark response (BMR), and Continuous Hitcall
- *Example on right: Bisphenol A*
  https://comptox.epa.gov/dashboard/chemical/invitrodb/DTXSID7020182

Application Programming Interfaces (APIs)
https://api-ccte.epa.gov/docs/bioactivity.html

Updated data view coming soon in 2024



- APIs provide data for various use cases, including research and applications with user interfaces
- Users can avoid large data downloads by accessing invitrodb programmatically via an API
- This is a great read-only solution for users who require more flexibility than the CCD can provide
- More integration with tcpl is coming soon and for additional documentation, check out the CCTE API Home Page: https://api-ccte.epa.gov/docs/index.html

ToxCast Data Downloads

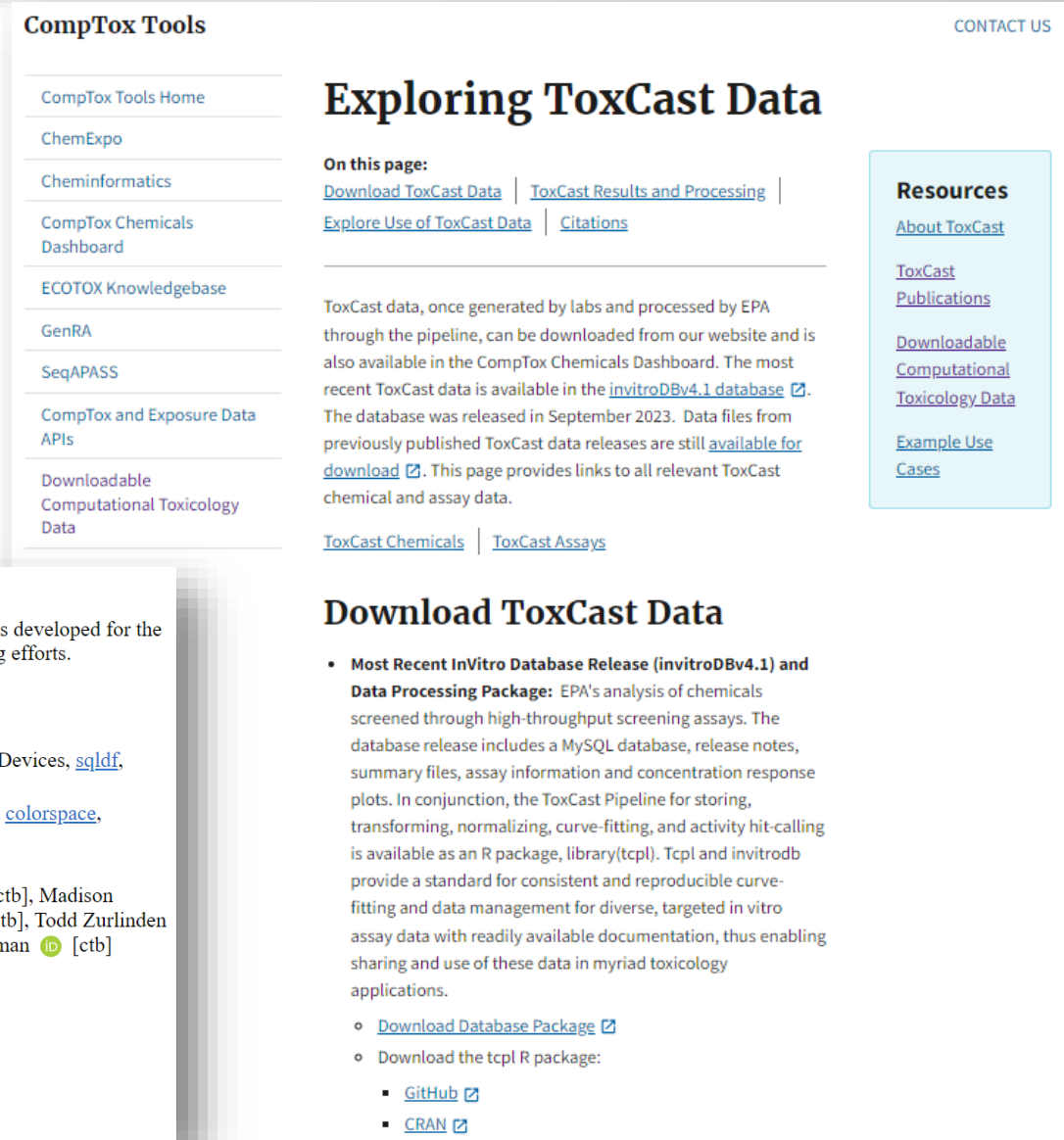
https://www.epa.gov/comptox-tools/exploring-toxcast-data

- Data downloads allow users to set up their own personal instance of the invitrodb MySQL database and interact with the data directly via the tcpl R package
- This is a preferred option for more customized or programmatic ToxCast data needs, or if users want to do their own data processing

tcpl: ToxCast Data Analysis Pipeline

A set of tools for processing and modeling high-throughput and high-content chemical screening data. The package was developed for the the chemical screening data generated by the US EPA ToxCast program, but can be used for diverse chemical screening efforts.

Version:	3.1.0
Depends:	R (≥ 3.5.0)
Imports:	data.table (≥ 1.9.4), DBI, RMariaDB, numDeriv, RColorBrewer, utils, stats, methods, graphics, grDevices, sqldf, dplyr, tidyr, plotly, tcplfit2, ggplot2, gridExtra, stringr
Suggests:	roxygen2, knitr, prettydoc, rmarkdown, htmlTable, testthat (≥ 3.0.0), reshape2, viridis, kableExtra, colorspace, magrittr, vdiffr
Published:	2023-10-06
Author:	Richard S Judson [ctb, ths], Dayne L Filer [aut], Jason Brown [cre], Sarah E Davidson-Fritz [ctb], Madison Feshuk [ctb], Lori Kolaczkowski [ctb], Kurt Dunham [ctb], Carter Thunes [ctb], Ashley Ko [ctb], Todd Zurlinden [ctb], Parth Kothiya [ctb], Woodrow R Setzer [ctb], Matthew T Martin [ctb, ths], Katie Paul Friedman [ctb]
Maintainer:	Jason Brown <brown.jason at epa.gov>
License:	MIT + file LICENSE
URL:	https://github.com/USEPA/CompTox-ToxCast-tcpl
NeedsCompilation:	no
Materials:	NEWS
CRAN checks:	tcpl results



Summary

- The ToxCast program makes targeted *in vitro* screening assay data publicly available for prioritization and hazard characterization.
- Data needs in next generation risk assessment necessitated software and database updates for consistent and reproducible curve-fitting and data management across screening efforts.
- Updates include additional models, bidirectional curve-fitting, and continuous hit calling.
- Annotation structure, fit categories, and cautionary flags on curve-fitting behavior were also modified.
- Curve-fitting updates resulted in small changes in activity hit calls and potency estimates but without a uniform trend.
- ToxCast data is accessible via the CompTox Chemicals Dashboard, APIs, and data downloads.

Thanks for listening!

Please reach out with questions

Madison Feshuk *Feshuk.Madison@epa.gov*
Jason Brown *brown.Jason@epa.gov*
Sarah Davidson-Fritz *DavidsonFritz.Sarah@epa.gov*
Katie Paul Friedman *Paul-Friedman.Katie@epa.gov*

Thank you to past contributors, collaborators, and our current ToxCast team:

Madison Feshuk
- *Lead, Pipelining and Curation*

Jason Brown
- *Lead, tcpl Development*

Sarah Davidson-Fritz
- *Lead, tcplfit2 Development and Statistics*

Carter Thunes
- *tcpl Development and Pipelining*
- ORAU-SSC

Ashley Ko
- *Pipelining and Curation*
- ORAU-SSC

Zhihui (Grace) Zhao
- *tcplfit2 Development*
- ORAU-SSC

Kelly Carstens
- *SME, DNT*

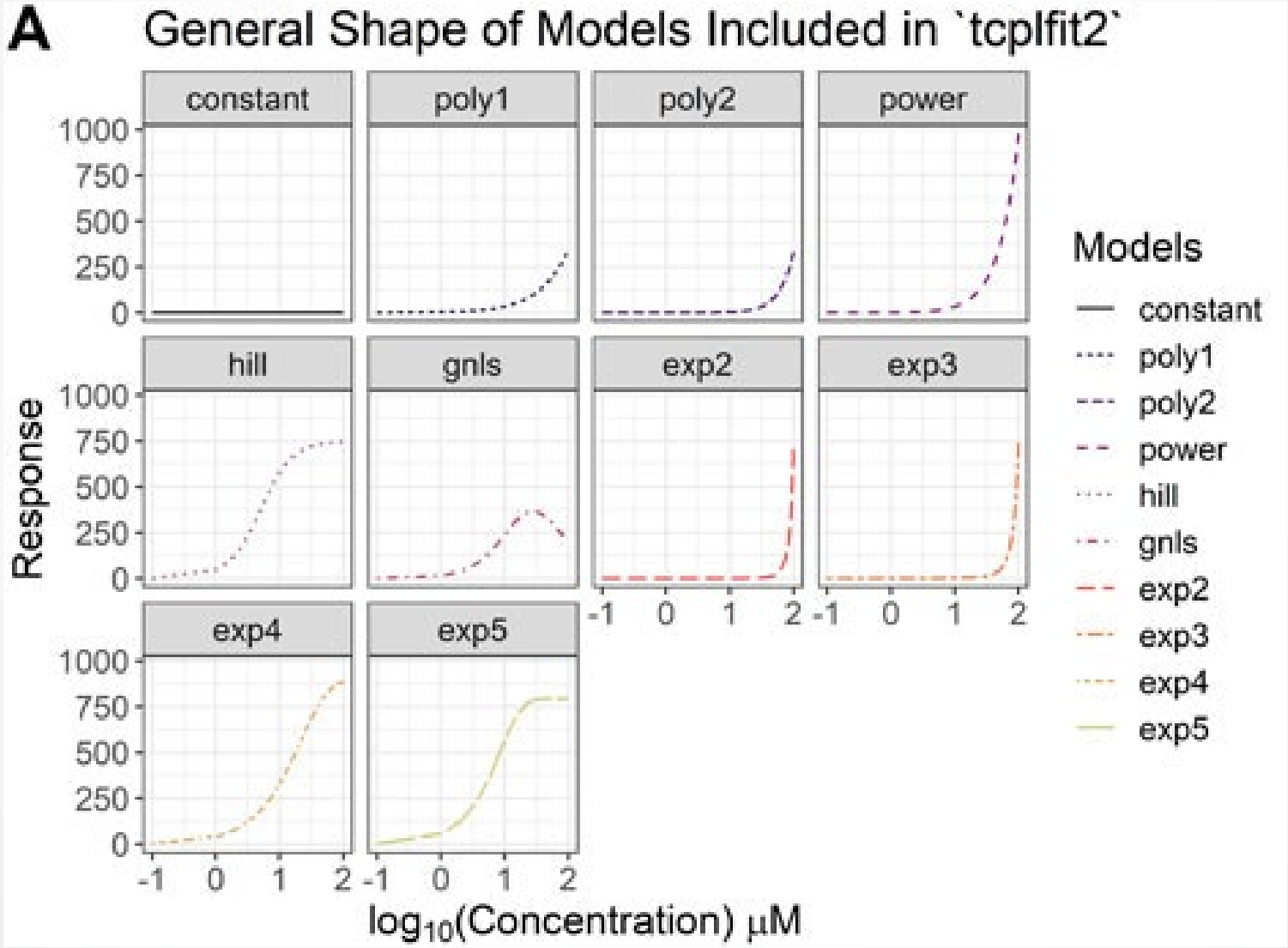
Katie Paul Friedman
- *SME, ToxCast Project Lead*

Richard Judson
- *SME*

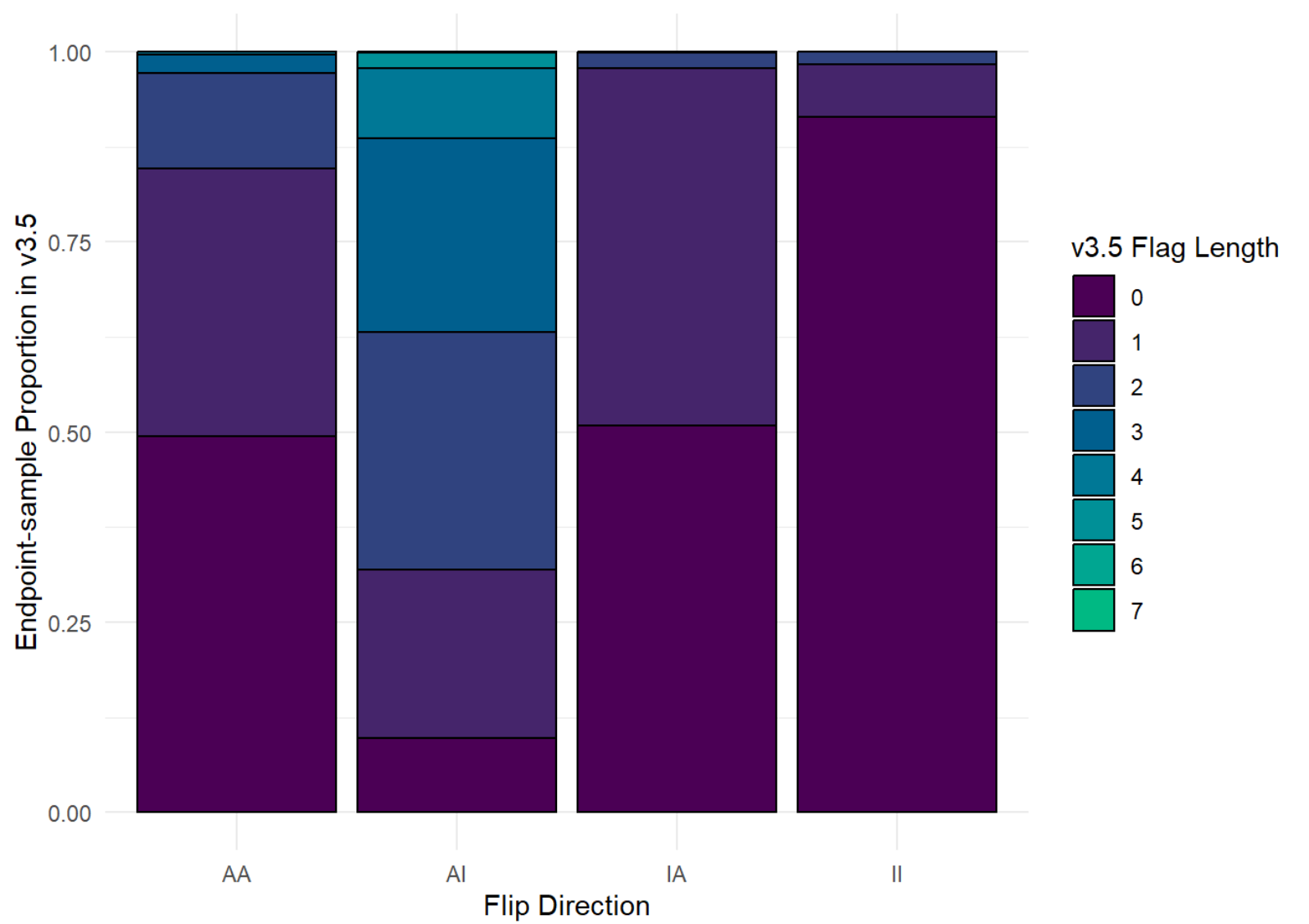
Colleen Elonen
- *Project Liaison*

Extra Slides

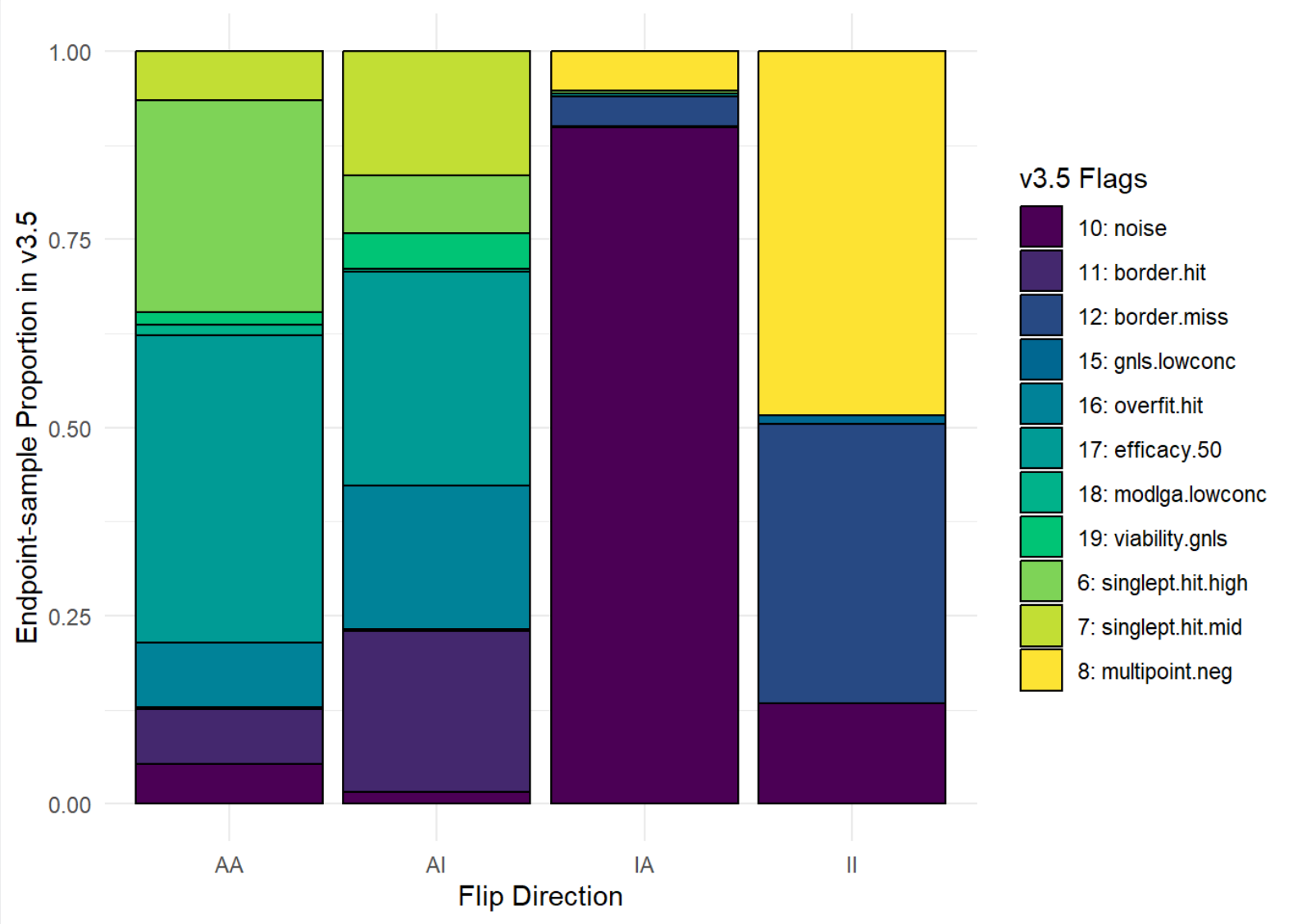
Shapes of Models



Flipped Hit Calls: Number of Flags



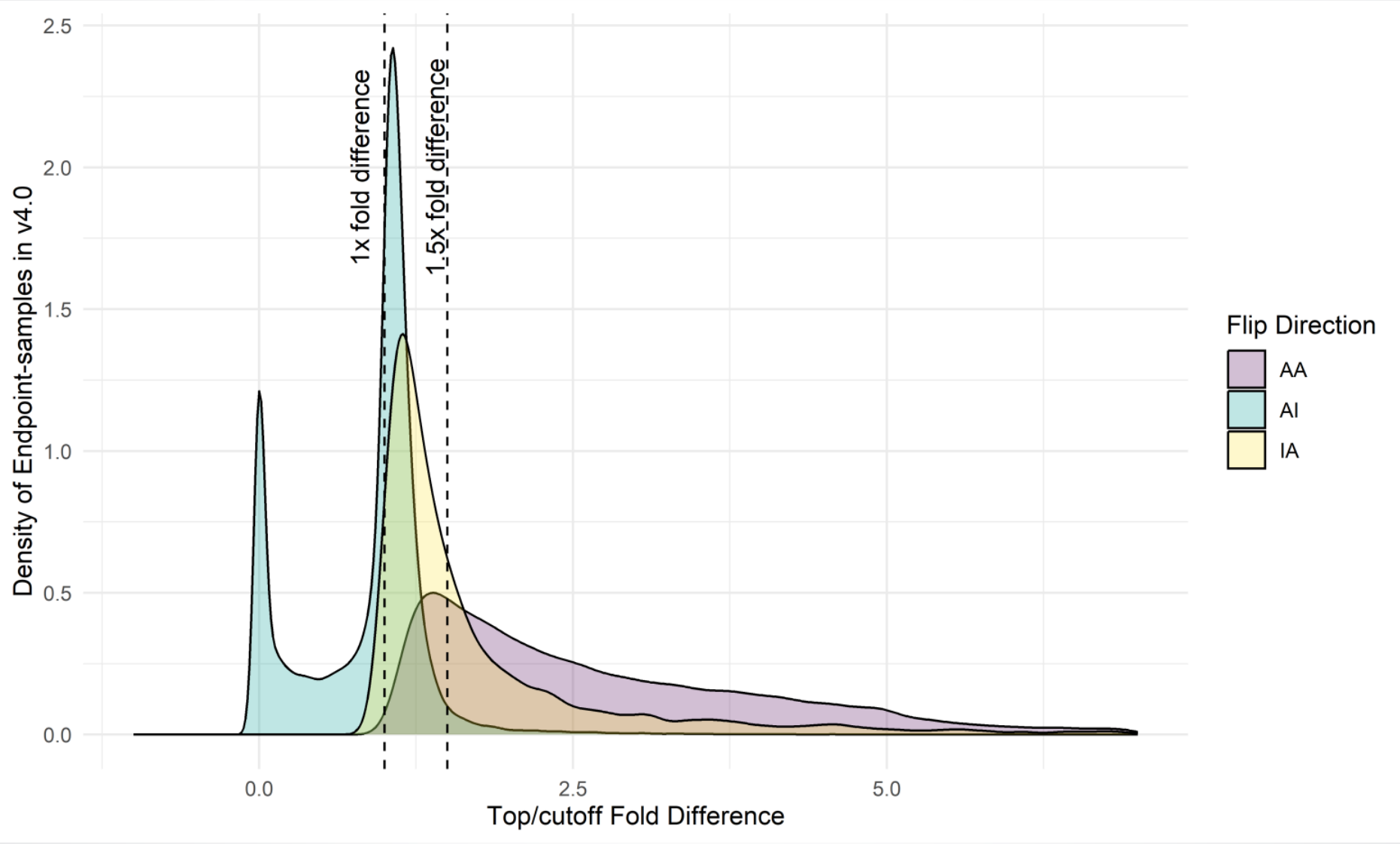
Flipped Hit Calls: Types of Flags



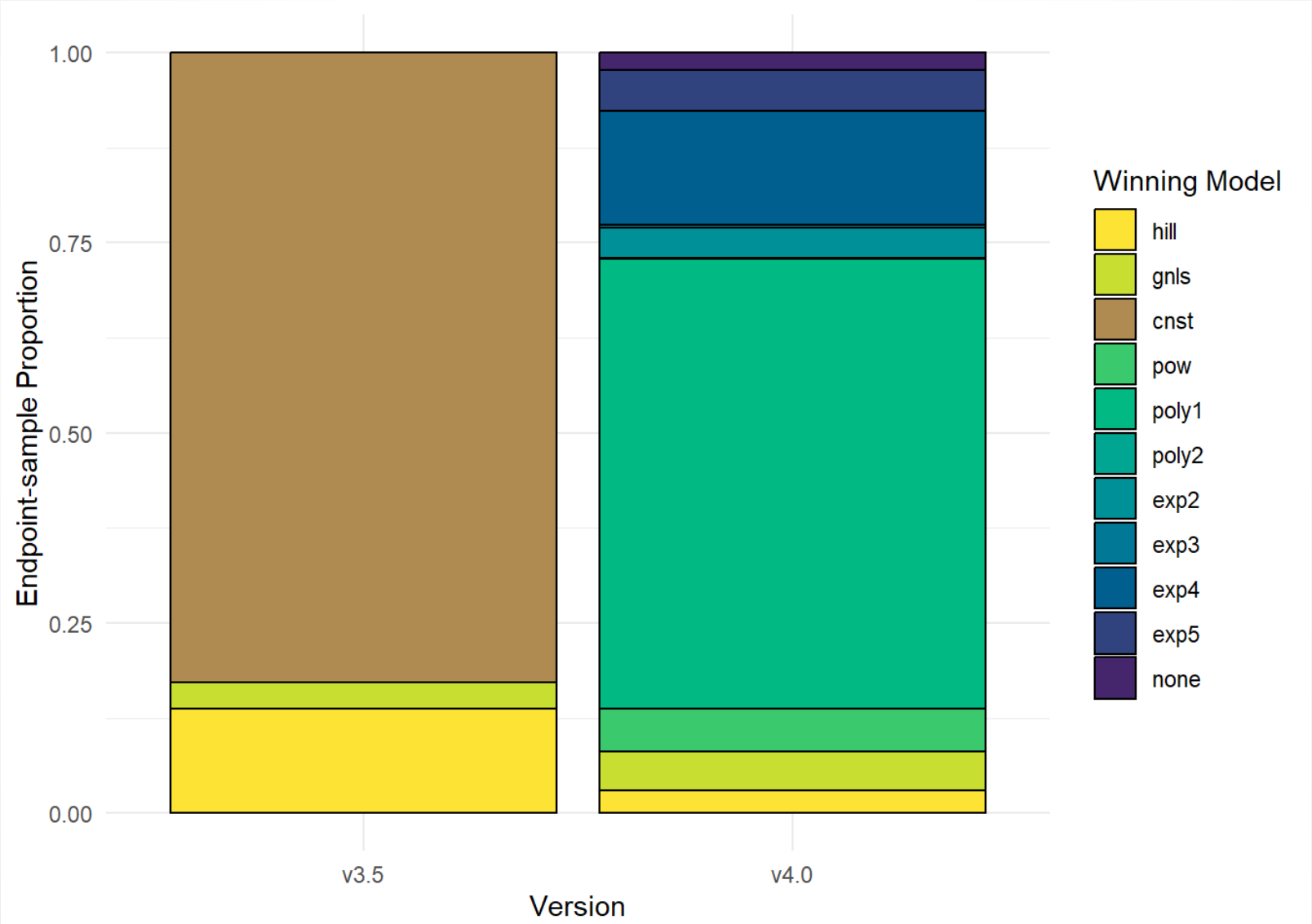
Flag Table Decoding

Level 6 Flag Name	Level 6 Flag Description
modl.directionality.fail	Model directionality is questionable as data points are split in positive and negative axis. tcplFit2 models assume data is zero-centered and the absolute response is increasing
low.nrep	Average number of replicates per conc is less than 2
low.nconc	Number of concentrations tested is less than 4
bmd.high	Bmd > ac50, indication of high baseline variability
singlept.hit.high	Only highest conc above baseline, active
singlept.hit.mid	Only one conc above baseline, active
multipoint.neg	Multiple points above baseline, inactive
gnls.lowconc	Complete gain-loss curve not within concentration range tested, as the “Gain” AC50 less than lowest concentration tested or the “Loss” AC50 greater than mean concentration tested
noise	Noisy data (rme>coff)
border	Borderline activity with top <= 1.2*coff or top >= 0.8*coff
efficacy.50	Less than 50% efficacy
ac50.lowconc	AC50 less than lowest concentration tested
viability.gnls	Cell viability assay fit with gain-loss winning model

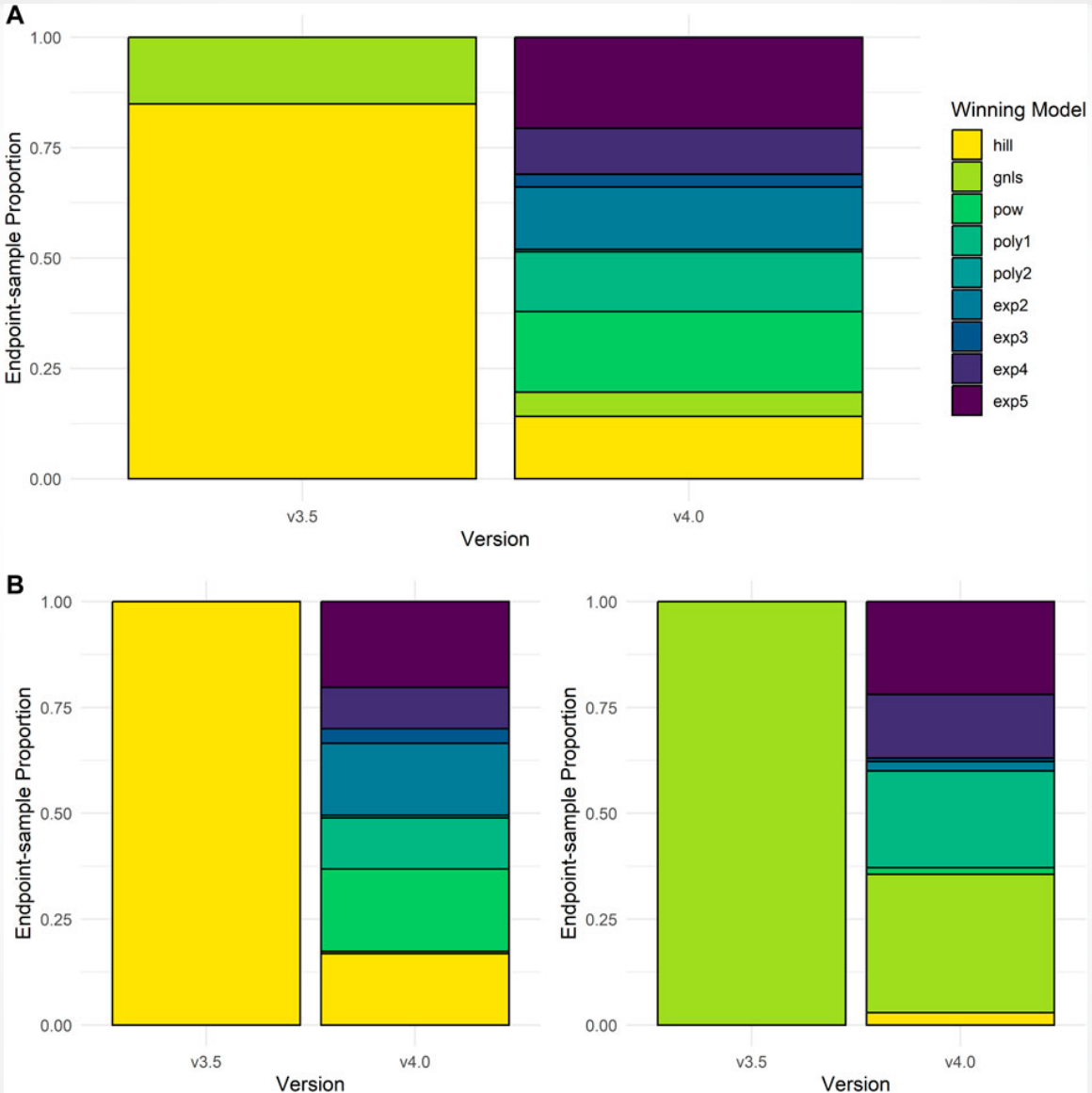
Top over Cutoff



Winning Model Selection



Winning Model Shifts: Actives



Cytotoxicity