

An Update on the ToxCast Data Pipeline: Features for Dataset Development

Katie Paul Friedman, PhD

Center for Computational Toxicology and Exposure, Office of Research
and Development, US Environmental Protection Agency

Research Triangle Park, NC

Email: paul-friedman.katie@epa.gov

*The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the
U.S. EPA*

Conflict of Interest Statement

The author declares no conflict of interest.

Overview of this presentation

- Introduction to ToxCast
- Briefly, what biology is covered by ToxCast?
- The ToxCast Data Pipeline (tcpl): How are ToxCast data managed and what are key definitions for use?
- Approaches to dataset development
 - Finding data by annotation
 - Filtering by curve-fit quality
 - Filtering for cytotoxicity/selectivity
- New features coming in tcpl version 3

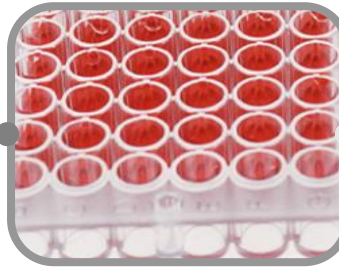
ToxCast / Tox21 Overall Strategy

- Identify targets or pathways linked to toxicity (AOP focus)
- Identify/develop high-throughput assays for these targets or pathways
- Develop predictive systems models
 - *in vitro/in silico* → *in vivo*
 - human focus
- Use predictive models:
 - Prioritize chemicals for targeted testing
 - Suggest / distinguish possible AOP / MOA for chemicals
- *High-throughput Exposure Predictions*
- *High-throughput Risk Assessments*

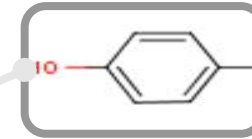
Hazard Predictions: High-Throughput Screening (HTS)



Robots



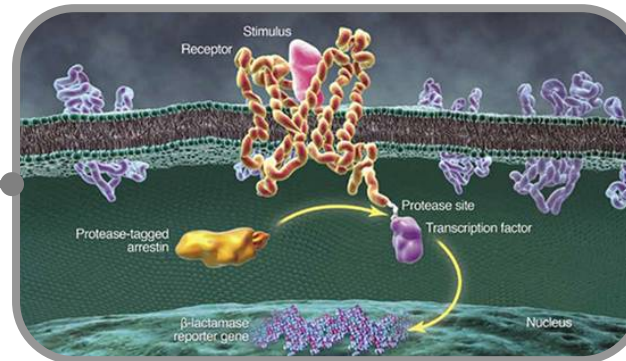
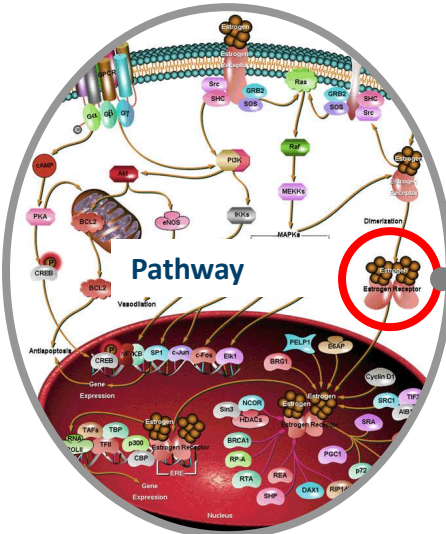
96-, 384-, 1536 Well Plates



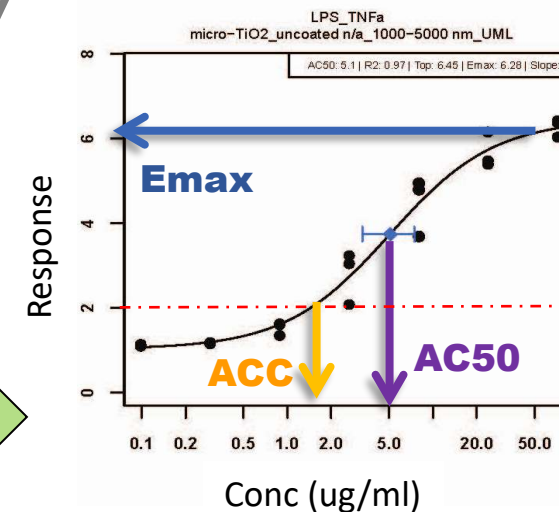
Chemical Exposure



Cell Population



Target Biology (e.g., Estrogen Receptor)



ToxCast begins with chemistry

Richard *et al.*, 2016

Chemical
Research in
Toxicology

Perspective
pubs.acs.org/cr

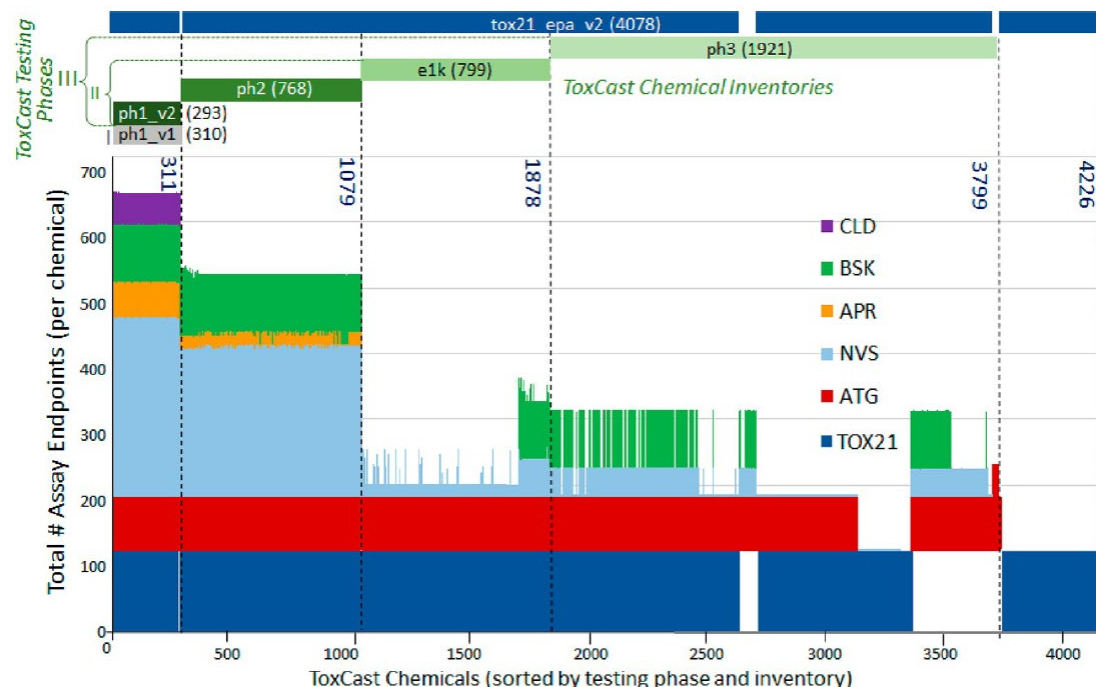
ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology

Ann M. Richard,^{*,†} Richard S. Judson,[†] Keith A. Houck,[†] Christopher M. Grulke,[†] Patra Volarath,[‡] Inthirany Thillainadarajah,[§] Chihae Yang,^{||,‡} James Rathman,^{‡,¶} Matthew T. Martin,[‡] John F. Wambaugh,[‡] Thomas B. Knudsen,[‡] Jayaram Kancherla,[‡] Kamel Mansouri,[‡] Grace Patlewicz,[‡] Antony J. Williams,[‡] Stephen B. Little,[‡] Kevin M. Crofton,[‡] and Russell S. Thomas[‡]

- Include pesticides, antimicrobials, contaminants, industrial, high production volume, lists with regulatory interest, FDA *in vivo* data sets, FDA food additives, fragrances, plasticizers, drugs
- ToxCast total substances: approaches 4,000
- Tox21 total substances: approaches 10,000

Chemical Research in Toxicology

Perspective



https://comptox.epa.gov/dashboard/chemical_lists/toxcast

What did we learn about bioactivity from screening large numbers of substances (100s to 10,000)?

- Assay performance could be defined
- New reference chemicals by target could be understood
- Integrated and predictive models could be built
- Prioritization based on bioactivity could be achieved

Screening large numbers of substances for bioactivity can illustrate trends, define domain of applicability, and better highlight strengths and weaknesses of the assays.

Bottom-line: building confidence

ToxCast contains heterogeneous data

Assay Sources

ACEA
Apredica
Attagene
BioSeek
CCTE/EPA ORD
CeeTox
CellzDirect
LifeTech Expression Analysis
NovaScreen (Perkin Elmer)
Odyssey Thera
Stemina
Tox21/NCATS
University Partners
Zebrafish: CCTE and Tanguay

Biological Response

cell proliferation and death
cell differentiation
Enzymatic activity
mitochondrial depolarization
protein stabilization
oxidative phosphorylation
reporter gene activation
gene expression (qNPA, RT-PCR)
receptor binding
receptor activity
Steroidogenesis
Metabolomic responses in stem cells

Target Family

response Element
transporter
cytokines
kinases
nuclear receptor
CYP450 / ADME
cholinesterase
phosphatases
proteases
XME metabolism
GPCRs
ion channels
ETC

Assay Design

viability reporter
morphology reporter
conformation reporter
enzyme reporter
membrane potential reporter
binding reporter
inducible reporter
ETC

Detection Technology

qNPA and ELISA
Fluorescence & Luminescence
Alamar Blue Reduction
Arraysan / Microscopy
Reporter gene activation
RT-PCR
Spectrophotometry
Radioactivity
HPLC and HPEC
TR-FRET

Readout Type

single
multiplexed
multiparametric

Cell Format

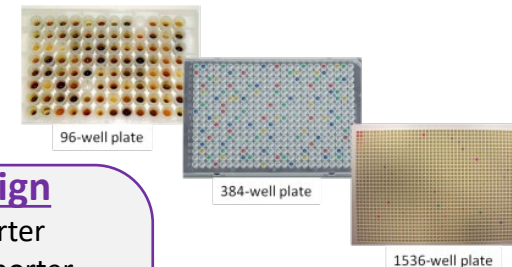
cell free
cell lines
primary cells
complex cultures
free embryos

Species

human
rat
mouse
zebrafish
sheep
boar
rabbit
cattle
guinea pig

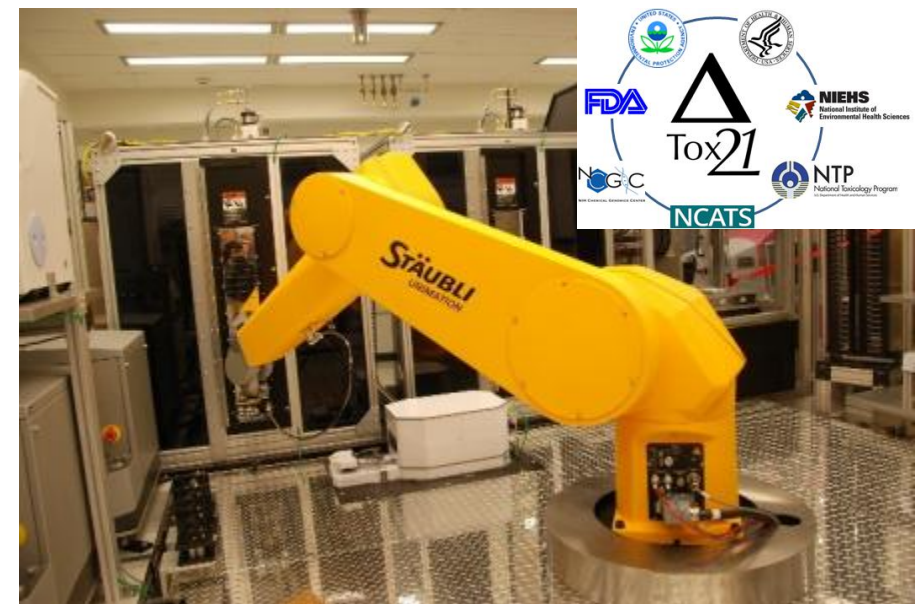
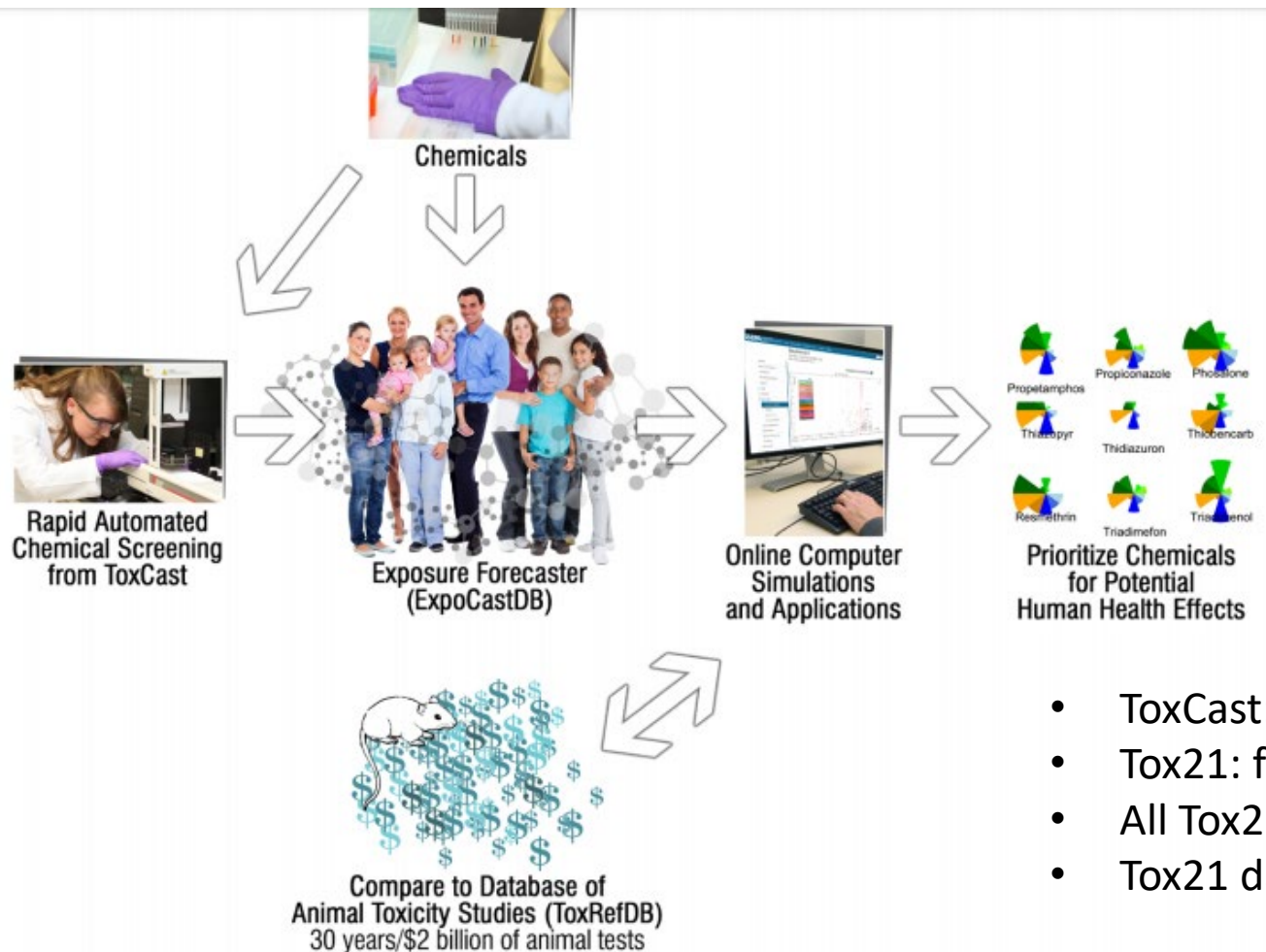
Tissue Source

Lung	Breast
Liver	Vascular
Skin	Kidney
Cervix	Testis
Uterus	Brain
Intestinal	Spleen
Bladder	Ovary
Pancreas	Prostate
Inflammatory	Bone



ToxCast and Tox21 have generated a lot of publicly available bioactivity data for hazard screening and prediction.

EPA's ToxCast program at a glance



Tox21 robot

- ToxCast: more assays, fewer chemicals, EPA-driven
- Tox21: fewer assays, all 1536, driven by consortium
- All Tox21 data are analyzed by multiple partners
- Tox21 data is available analyzed in the ToxCast Data Pipeline

ToxCast covers a lot of biology but not all; and, ToxCast is growing over time.

Invitrodb version 3.3 (released August 2020) contained 17 different assay sources, covering (at least) 491 unique gene-related targets with 1600 unique assay endpoints. Varying amounts of data are available for 9949 unique substances.

Assay source	Long name	Truncated assay source description	Some rough notes on the biology covered
ACEA	ACEA Biosciences	real-time, label-free, cell growth assay system based on a microelectronic impedance readout	Endocrine (ER-induced proliferation)
APR	Apredica	CellCiphr High Content Imaging system	Hepatic cells (HepG2)
ATG	Attagene	multiplexed pathway profiling platform	Nuclear receptor and stress response profile
BSK	Bioseek	BioMAP system providing uniquely informative biological activity profiles in complex human primary co-culture systems	Immune/inflammation responses
NVS	Novascreen	large diverse suite of cell-free binding and biochemical assays.	Receptor binding; transporter protein binding; ion channels; enzyme inhibition; many targets
OT	Odyssey Thera	novel protein:protein interaction assays using protein-fragment complementation technology	Endocrine (ER and AR)
TOX21	Tox21/NCGC	Tox21 is an interagency agreement between the NIH, NTP, FDA and EPA. NIH Chemical Genomics Center (NCGC) is the primary screening facility running ultra high-throughput screening assays across a large interagency-developed chemical library	Many – with many nuclear receptors
CEETOX	Ceetox/OpAns	HT-H295R assay	Endocrine (steroidogenesis)
CLD	CellzDirect	Formerly CellzDirect, this Contract Research Organization (CRO) is now part of the Invitrogen brand of Thermo Fisher providing cell-based in vitro assay screening services using primary hepatocytes.	Liver (Phase I/Phase II/ Phase III expression)
NHEERL_PADILLA	NHEERL Padilla Lab	The Padilla laboratory at the EPA National Health and Environmental Effects Research Laboratory focuses on the development and screening of zebrafish assays.	Zebrafish terata
NCCT	NCCT Simmons Lab	The Simmons Lab at the EPA National Center for Computational Toxicology focuses on developing and implementing in vitro methods to identify potential environmental toxicants.	Endocrine (thyroid - thyroperoxidase inhibition)
TANGUAY	Tanguay Lab	The Tanguay Lab, based at the Oregon State University Sinnhuber Aquatic Research Laboratory, uses zebrafish as a systems toxicology model.	Zebrafish terata/phenotypes
NHEERL_NIS	NHEERL Stoker & Laws	The Stoker and Laws laboratories at the EPA National Health and Environmental Effects Research Laboratory work on the development and implementation of high-throughput assays, particularly related to the sodium-iodide cotransporter (NIS).	Endocrine (thyroid - NIS inhibition)
UPITT	University of Pittsburgh	The Johnston Lab at the University of Pittsburgh ran androgen receptor nuclear translocation assays under a Material Transfer Agreement (MTA) for the ToxCast Phase 1, Phase 2, and E1K chemicals.	Endocrine (AR related)

With each release, more assay endpoints and more chemical x endpoint data are released

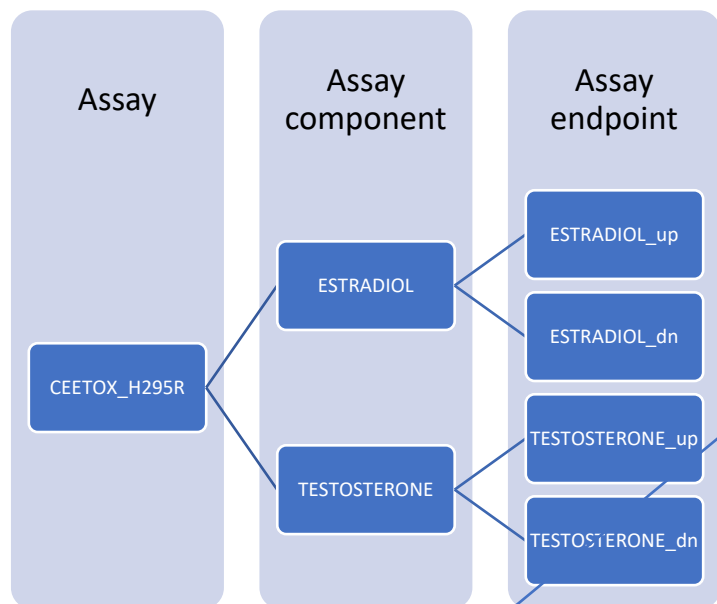
Invitrodb version 3.3 (released August 2020) contained 17 different assay sources, covering (at least) 491 unique gene-related targets with 1600 unique assay endpoints. Varying amounts of data are available for 9949 unique substances.

These assay endpoints were notable additions in invitrodb version 3.3.

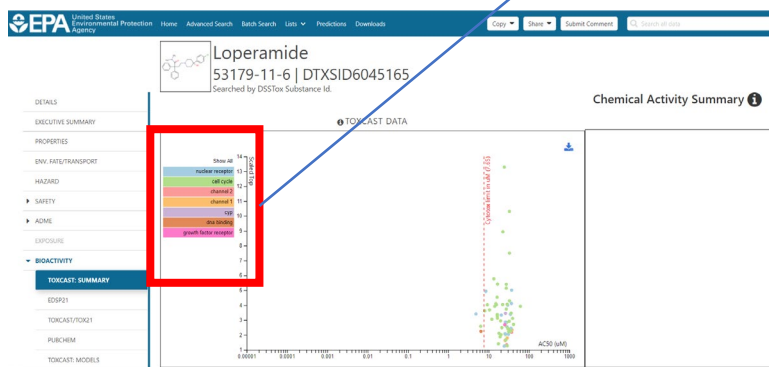
Assay source	Long name	Truncated assay source description	Some rough notes on the biology covered
NCCT_MITO	NCCT (now Center for Computational Toxicology and Exposure) Mitochondrial toxicity	Respirometric assay that measure mitochondrial function in HepG2 cells	Multiple assay endpoints to evaluate mitochondrial function https://doi.org/10.1093/toxsci/kfaa059 .
NHEERL_MED	NHEERL Mid-Continent Ecology Division	The EPA Mid-Continent Ecology Division of the National Health and Environmental Effects Research Laboratory screened the ToxCast Phase 1 chemical library for hDIO1 (deiodinase 1) inhibition as part of an ecotoxicology effort.	Endocrine (thyroid – hDIO1,2,3 inhibition) https://doi.org/10.1093/toxsci/kfy302
STM	Stemina	Stem cell-based metabolomic indicator of developmental toxicity for screening.	Developmental toxicity screening – multiple assay endpoints https://doi.org/10.1093/toxsci/kfaa014
LTEA	Life Tech Expression Analysis	Gene expression measured in HepaRG cells following 48 hr exposure	Liver toxicity model via transcription factor regulated-metabolism and markers of oxidative/cell stress; multiple assay endpoints

Learning more about the assay endpoints and biology

Example assay annotation hierarchy




- Many assay endpoints are mapped to a gene, if applicable
- Assay endpoints now cover 1398 unique gene targets in invitrodb version 3.3, in addition to other processes
- Intended target family is one way to understand biological target (incomplete list here):
 - Apolipoprotein
 - Apoptosis
 - Background measurement
 - Catalase
 - Cell adhesion
 - Cell cycle
 - Cell morphology
 - CYP
 - Cytokine
 - Deiodinase
 - DNA binding
 - Esterase
 - Filaments
 - GPCR
 - Growth factor
 - Histones
 - Hydrolase
 - Ion channel
 - Kinase
 - Ligase
 - Lyase
 - Malformation (zebrafish)
 - Membrane protein
 - Metabolite (Stemina metabolomics)
 - Mitochondria
 - Methyltransferase
 - microRNA
 - Mutagenicity response
 - Nuclear receptor
 - Oxidoreductase
 - Phosphatase
 - Protease/inhibitor
 - Steroid hormone
 - Transferase
 - Transporter



https://comptox.epa.gov/dashboard/assay_endpoints/


Download summary information here: <https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data>

More information about assay endpoints

 **United States Environmental Protection Agency**



HomeAdvanced SearchBatch SearchLists ▼PredictionsDownloads

Share 🔽

 Search all data

Download 🔽

10 ▼

Assay Component Endpoint Name	Details	Multi-Targeted	Active
ACEA_ER_80hr		456 / 3024	
APR_HepG2_CellCycleArrest_1h_dn		3 / 310	

All Chemicals in Assay Endpoint: [ACEA_ER_80hr](#)

Excel

AnnotationsCitationstcpl ProcessingReagentsAOPs

Annotations

Citations

tcpl Processing

Reagents

AOPs

PMID	url	Title	Author	Citation	doi
1 16481145	PubMed URL	Microelectronic cell sensor assay for detection of cytotoxicity and prediction of acute toxicity	Xing JZ, Zhu L, Gabos S, Xie L	Xing JZ, Zhu L, Gabos S, Xie L. Microelectronic cell sensor assay for detection of cytotoxicity and prediction of acute toxicity. Toxicol In Vitro. 2006 Sep;20(6):995-1004. Epub 2006 Feb 14. PubMed PMID: 16481145.	
2 23682706	PubMed URL	Real-time growth kinetics measuring hormone mimicry for ToxCast chemicals in T-47D human ductal carcinoma cells	Rotroff DM, Dix DJ, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Reif DM, Richard AM, Sipes NS, Abassi YA, Jin C, Stampfl M, Judson RS	Rotroff DM, Dix DJ, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Reif DM, Richard AM, Sipes NS, Abassi YA, Jin C, Stampfl M, Judson RS. Real-time growth kinetics measuring hormone mimicry for ToxCast chemicals in T-47D human ductal carcinoma cells. Chem Res Toxicol. 2013 Jul 15;26(7):1097-107. doi:10.1021/tx400117y. Epub 2013 Jun 10. PubMed PMID: 23682706.	doi: 10.1021/tx400117y

Assay Component Endpoint Name

ACEA_ER_80hr

Assay Component Endpoint Desc

Data from the was analyzed growth report they relate to produced mu targets, this a

Assay Function Type

signaling

Normalized Data Type

percent_activ

Analysis Direction

positive

Burst Assay

false

Key Positive Control

17b-estradiol

Signal Direction

gain

Intended Target Type

pathway

Intended Target Type Sub

pathway-specified

Intended Target Family

nuclear receptor

Intended Target Family Sub

steroidal

Assay Component Name

ACEA_ER_80hr

Assay Component Desc

ACEA_ER_80hr, is one of two assay component(s) measured or calculated from the ACEA_ER assay. It is designed to make measurements of real-time cell-growth kinetics, a form of growth reporter, as detected with electrical impedance signals by Real-

How are ToxCast/Tox21 data managed and what are the key data definitions for use?

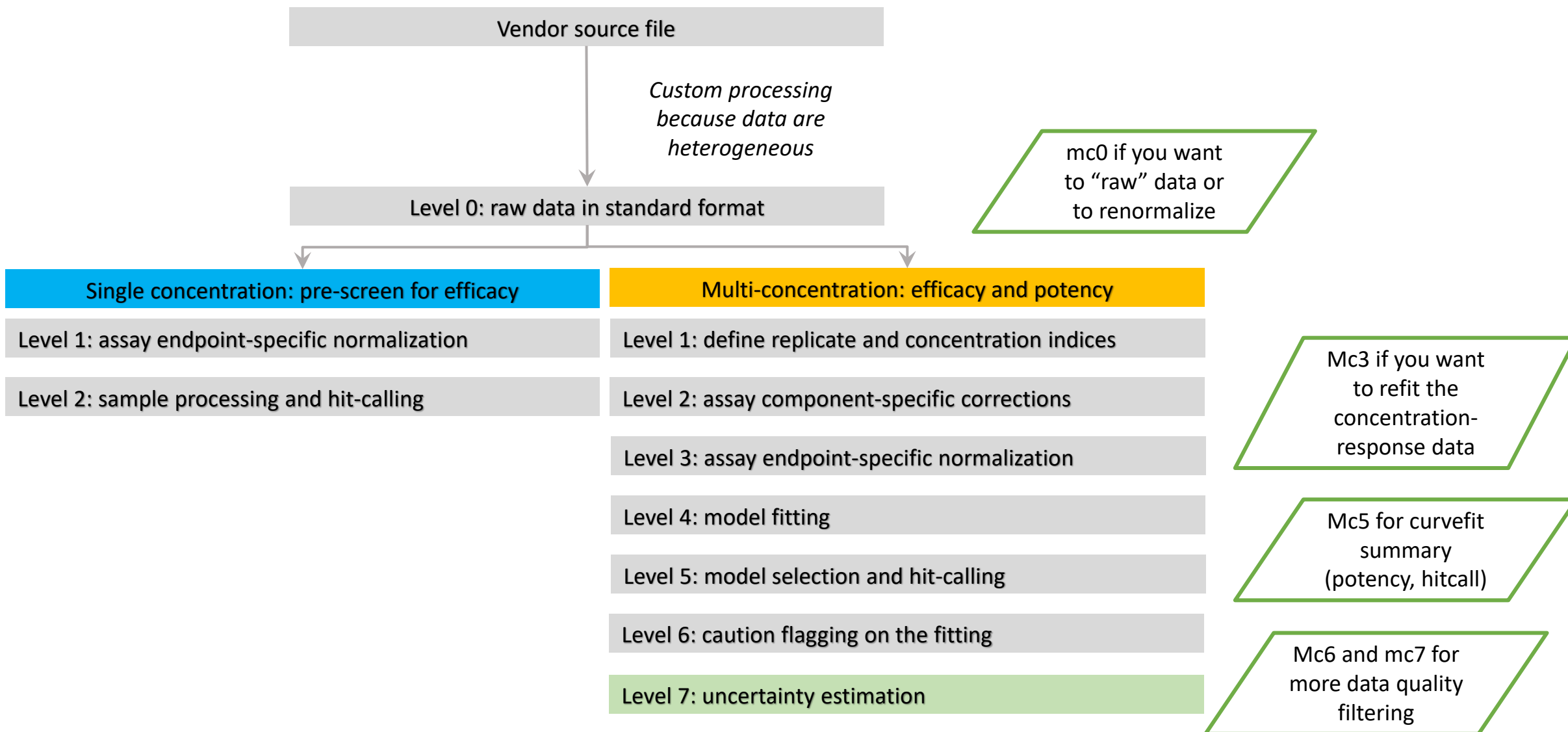
Summary information, datasets, and the full database (invitrodb version 3.3 August 2020 release) are available here:

<https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data>

ToxCast Pipeline (tcpl) Overview

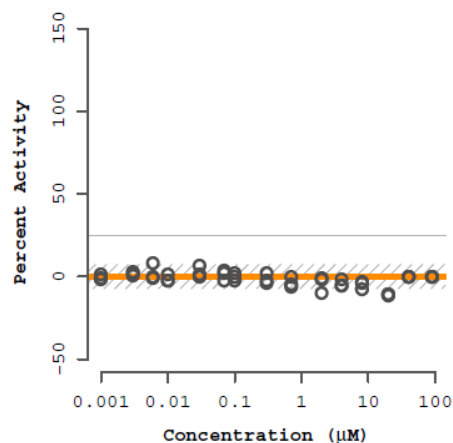
- Open source R software
- “Raw” source data at level 0
- Systematic but flexible analysis, with documentation of the methods applied to transform, normalize, curve-fit, call hits, apply cautions, etc.
- Storage of data at “levels” to standardize for any future analysis and make heterogeneous data into a homogeneous form
- Use combination of statistics (x-MAD, AIC) and biology-based efficacy cutoffs
- Points of departure (e.g. AC10, ACC, AC50) are included
- System of “caution flags” has been developed (continues to evolve)
- Beta uncertainty information based on bootstrap resampling

ToxCast: high-throughput bioactivity information

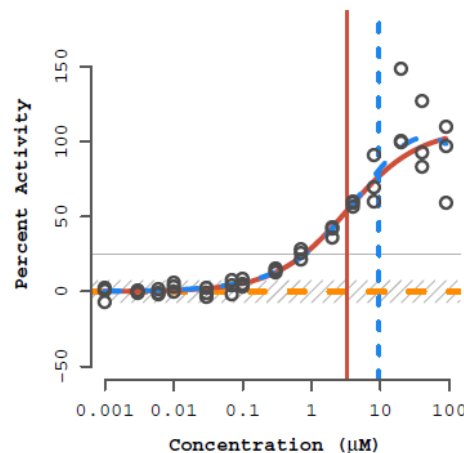


Pipeline Overview: Curve Fits

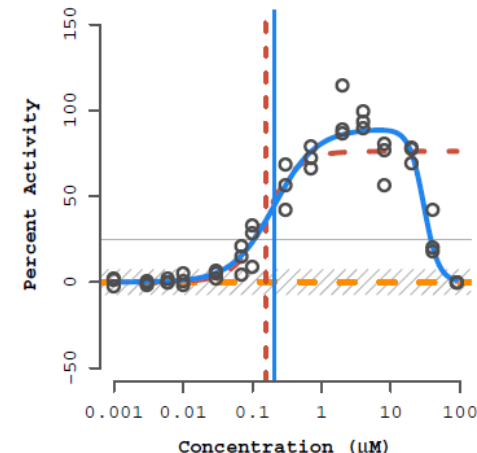
CONSTANT (cnst)



HILL (hill)



GAIN-LOSS (gnls)



Model winner determined by lowest AIC

Tcpl is on CRAN and GitHub with 1-2 updates a year

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: 2.0.2
Depends: R ($\geq 3.2.0$)
Imports: [data.table](#) ($\geq 1.9.4$), [DBI](#), [RMySQL](#), [numDeriv](#), [RColorBrewer](#), [utils](#), [stats](#), [graphics](#), [grDevices](#), [sqldf](#)
Suggests: [roxygen2](#), [knitr](#), [prettydoc](#), [rmarkdown](#), [htmlTable](#)
Published: 2019-07-26
Author: Richard S Judson [cre, ths], Dayne L Filer [aut], Jason Brown [ctb], Todd Zurlinden [ctb], Parth Kothiyi [ctb], Woodrow R Setzer [ctb], Matthew T Martin [ctb, ths], Katie Paul Friedman [ctb]
Maintainer: Richard S Judson <Judson.Richard@epa.gov>
License: [GPL-2](#)
URL: <https://github.com/USEPA/CompTox-ToxCast-tcpl>
NeedsCompilation: no
Materials: [NEWS](#)
CRAN checks: [tcpl results](#)

Vignettes on CRAN and peer-reviewed work

Bioinformatics, 33(4), 2017, 618–620
doi: 10.1093/bioinformatics/btw680
Advance Access Publication Date: 22 November 2016
Applications Note



Data and text mining

tcpl: the ToxCast pipeline for high-throughput screening data

Dayne L. Filer¹, Parth Kothiyi¹, R. Woodrow Setzer², Richard S. Judson² and Matthew T. Martin^{2,*}

Key ToxCast vocabulary for using these data

Key vocabulary	Full description	Derivation	Use
AC50	50% activity concentration, often represented as log10-AC50 (micromolar units)	A stable point on the curve that is 50% of the <u>maximal fitted response</u>	Potency estimate
ACC	Activity concentration at the cutoff, often represented as log10-ACC (micromolar units)	Similar to a benchmark dose; variable efficacy across heterogeneous assays	Potency estimate
HITC	Hitcall: -1, 0, 1	Qualitative activity determination; hitc=-1 not enough data to fit; hitc=0 negative because model top does not exceed the coff and/or the winning model is constant; hitc=1 positive	Binary activity – pretty incomplete picture (think borderline efficacy)
COFF	Efficacy “cut-off”	Statistical or biology-based cut-off for a positive; assay endpoint-dependent	Determines positive/negative hitcall
BMAD	Baseline median absolute deviation	Median absolute deviation of data that approximate assay “baseline;” can be lowest two concentrations in the index (by plate), or can be DMSO or vehicle wells	3*BMAD is a common way to bound the “noise” in the assay baseline so that signal can be distinguished from noise
Flags	Caution flags on curve-fitting (from level 6)	Lots of different specific flags from “borderline activity” to “noisy fit”	Not all curve fits with flags are bad; some flags worse than others; >= 3 flags tend to indicate low quality curves
Model, winning model	Curve-fitting models (e.g., Hill, gain-loss, constant)	The winning model has the lowest AIC	the winning model determines the potency estimates reported

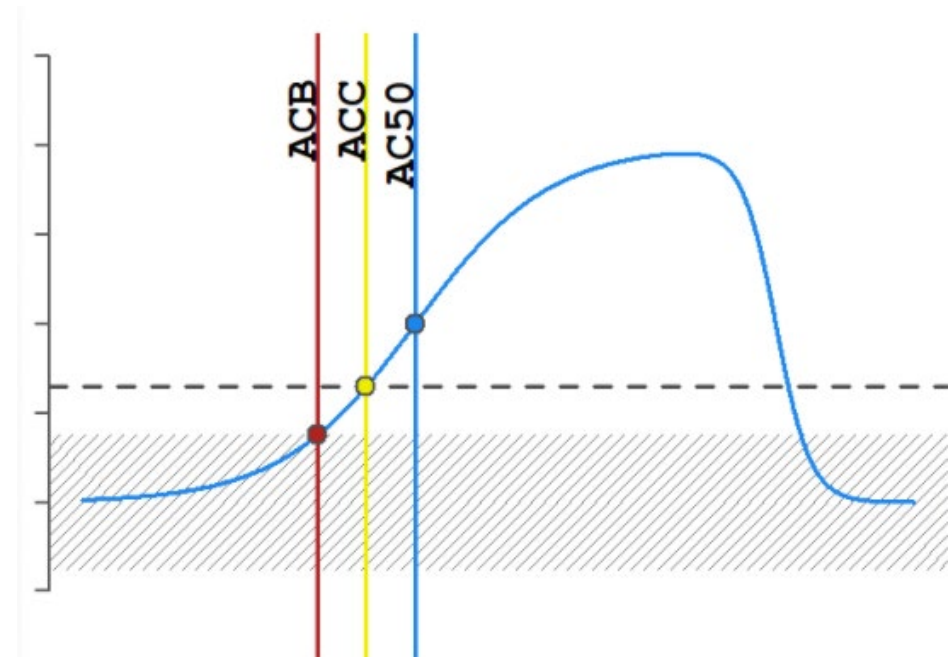
Additional definitions for mc5 table and beyond

Table 15: Fields in mc5 table.

Field	Description
m5id	Level 5 ID
m4id	Level 4 ID
aeid	Assay endpoint ID
modl	Winning model: "cnst", "hill", or "gnls"
hitc	Hit-/activity-call, 1 if active, 0 if inactive, -1 if cannot determine
fitc	Fit category
coff	Efficacy cutoff value
actp	Activity probability ($1 - \text{const_prob}$)
modl_er	Scale term for the winning model
modl_tp	Top asymptote for the winning model
modl_ga	Gain AC_{50} for the winning model
modl_gw	Gain Hill coefficient for the winning model
modl_la	Loss AC_{50} for the winning model
modl_lw	Loss Hill coefficient for the winning model
modl_prob	Probability for the winning model
modl_rmse	RMSE for the winning model
modl_acc	Activity concentration at cutoff for the winning model
modl_acb	Activity concentration at baseline for the winning model
modl_ac10	AC_{10} for the winning model

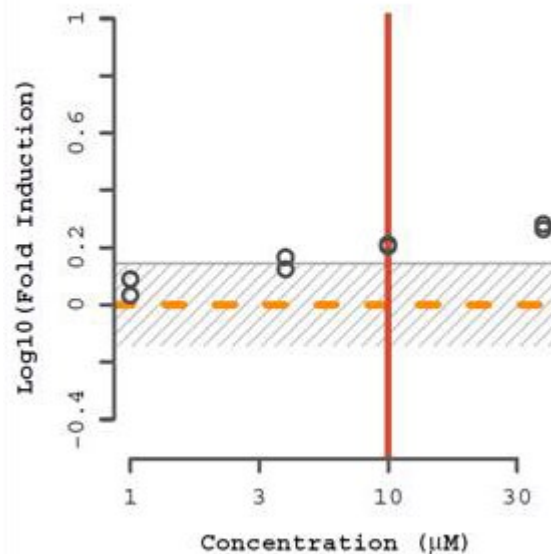
https://cran.r-project.org/web/packages/tcpl/vignettes/Introduction_Appendices.html

Has all of the definitions for every field in all of the single and multi-concentration tables of invitrodb

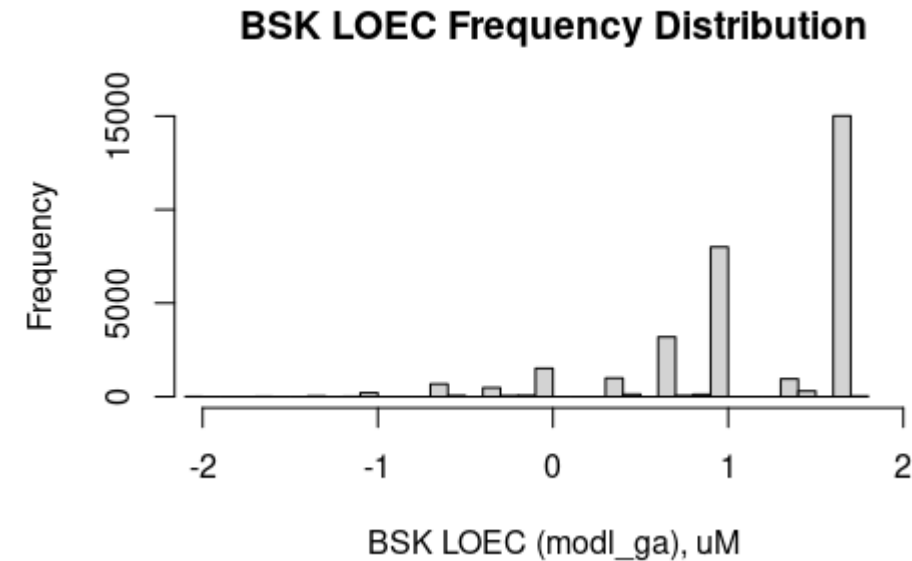


Adaptability: special cases

- BioSeek (BSK) data most commonly were obtained at 4 concentrations in duplicate, with limited dynamic range.
- The biological signals seemed relevant, but previously a subset of curve-fits (particularly gain-loss) drove very low AC50 values, at times lower than the concentration range screened, for effects of low magnitude. These were not very informative quantitatively.
- Lowest observable effect concentration method was written and implemented, and now BSK modl_ga column is populated with the LOEC value



First conc > coff; coff varies by endpoint and is the greater of 3BMAD or log10(1.2) change



The distribution is somewhat discretized because of the LOEC method; among hitc==1, LOECs at about 1, 4, 10, and 40 μM

Approaches to dataset development for further modeling

See the tcpl CRAN vignettes:

https://cran.r-project.org/web/packages/tcpl/vignettes/Data_retrieval.html

https://cran.r-project.org/web/packages/tcpl/vignettes/Data_processing.html

Finding assay data of interest with tcpl and R::RMySQL

- Using tcpl() functions alone
 - tcplLoadAsid(), tcplLoadAcid(), tcplLoadAeid() are helper functions to find assay sources, components, and endpoints; can group on any number of terms in the returned data.table.

```
> tcplLoadAsid()
```

	asid	asnm
1:	1	ACEA
2:	2	
3:	3	
4:	4	
5:	5	

```
> tcplLoadAcid(fld='asid',val=1)
```

	asid	acid	acnm
1:	1	1	ACEA_ER_80hr
2:	1	1804	ACEA_ER_AUC_viability
3:	1	1802	
4:	1	1829	
5:	1	1831	ACE
6:	1	1830	

```
> tcplLoadAeid(fld='asid',val=1)
```

	asid	aeid	aenm
1:	1	2	ACEA_ER_80hr
2:	1	1852	ACEA_ER_AUC_viability
3:	1	1850	ACEA_AR_agonist_AUC_viability
4:	1	1855	ACEA_AR_agonist_80hr
5:	1	1857	ACEA_AR_antagonist_AUC_viability
6:	1	1856	ACEA_AR_antagonist_80hr

Finding assay data of interest with tcpl and R::RMySQL

- Using R::RMySQL or summary files from the release .zip:
 - Use the mapped gene (as a surrogate for biology)
 - Use the intended target family or subfamily

```
assay <- dbGetQuery(con, "SELECT * FROM invitrodb.assay;") %>% data.table()
assay.component <- dbGetQuery(con, "SELECT * FROM invitrodb.assay_component;") %>% data.table()
assay.component.endpoint <- dbGetQuery(con,
                                         "SELECT * FROM invitrodb.assay_component_endpoint;") %>% data.table()
```

Alternatively, the *assay_annotation_information_invitrodb_v3_3.xlsx* contains this information in the INVITRODB SUMMARY file download from <https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data>

A	B	C	M	N	O	P
aeid	acid	assay_component_endpoint_name	intended_target_type	intended_target_type_sub	intended_target_family	intended_target_family_sub
2	1	ACEA_ER_80hr	pathway	pathway-specified	nuclear receptor	steroidal
3	2	APR_HepG2_CellCycleArrest_1h_dn	pathway	pathway-specified	cell cycle	proliferation

Finding assay data of interest with tcpl and R::RMySQL

- Using R::RMySQL or summary files from the release .zip:
 - Use the mapped gene (as a surrogate for biology)
 - Use the intended target family or subfamily

```
gene.target <- dbGetQuery(con, "SELECT * FROM invitrodb.gene  
                                INNER JOIN invitrodb.intended_target ON  
                                |gene.gene_id = intended_target.target_id;") %>% data.table()
```

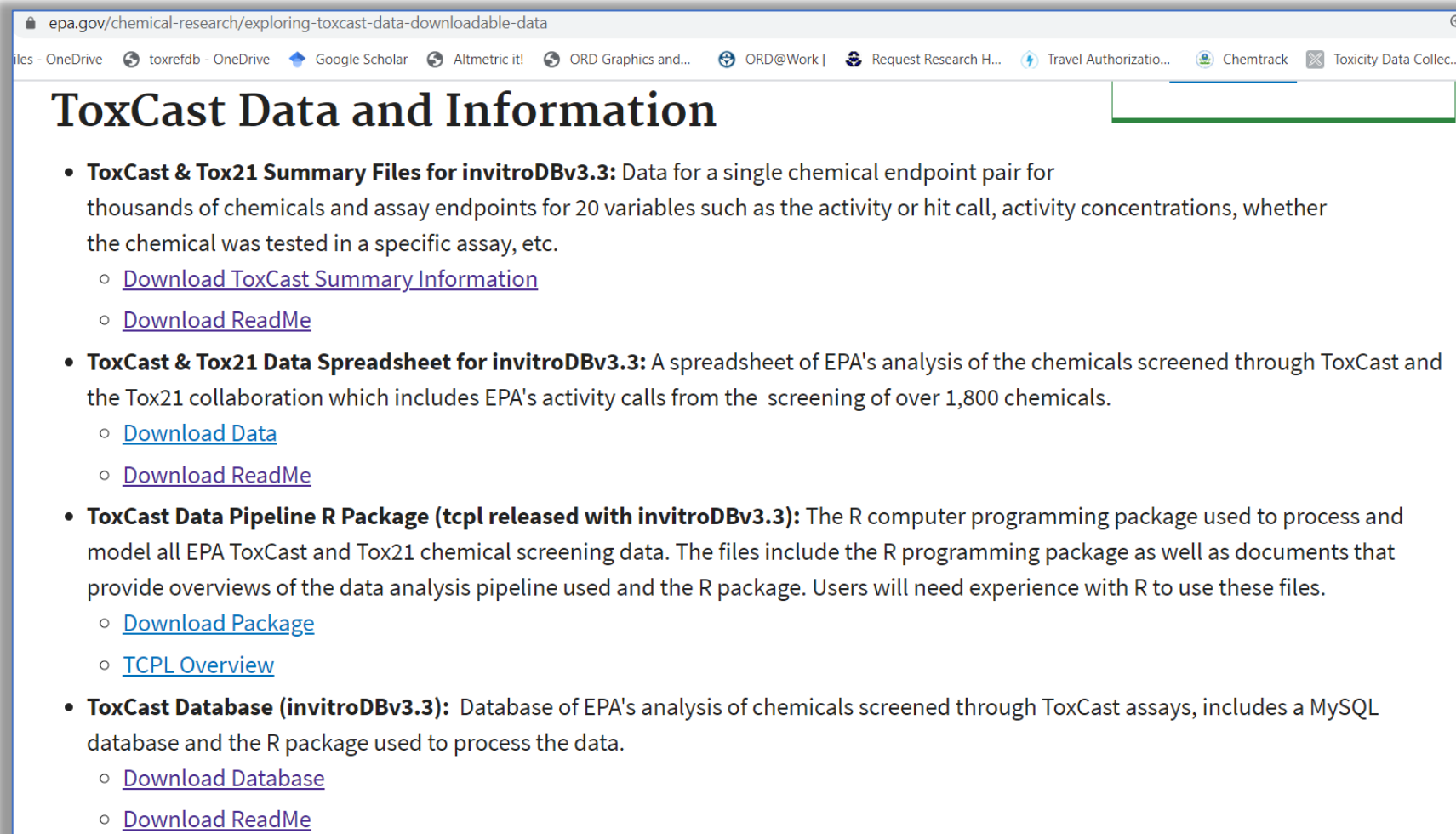
Alternatively, the *assay_annotation_information_invitrodb_v3_3.xlsx* contains this information in the INVITRODB SUMMARY file download from <https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data>

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
gene_id	entrez_ge	official_full_name	gene_name	official_sy	gene_sym	descriptio	uniprot_a	organism	track_stat	aeid	target_id	source	aenm		
117	2099	estrogen receptor 1	estrogen receptor 1	ESR1	ESR1	#N/A	P03372	1	live	2	117	gene	ACEA_ER_80hr		
329	7157	tumor protein p53	tumor protein p53	TP53	TP53	#N/A	P04637	1	live	19	329	gene	APR_HepG2_p53Act_1h_dn		
329	7157	tumor protein p53	tumor protein p53	TP53	TP53	#N/A	P04637	1	live	20	329	gene	APR_HepG2_p53Act_1h_up		

New release file format

<https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data>

- INVITRODB_V3_3_SUMMARY.zip from our downloads page now contains a number of helpful flat files, including all mc5 and mc6 data as .Rdata and .xlsx by vendor source name.
- Assay annotation information is now more helpful and resembles the database itself.
- No more putting together lots of .csv files to find hitcalls – just load the flat files with mc5+mc6.
- See: DB_release_README_SUMMARY.pdf and INVITRODB_V3_3_SUMMARY.zip for a full listing of files and description.



The screenshot shows the EPA website page titled "ToxCast Data and Information". The page lists four main categories of data available for download:

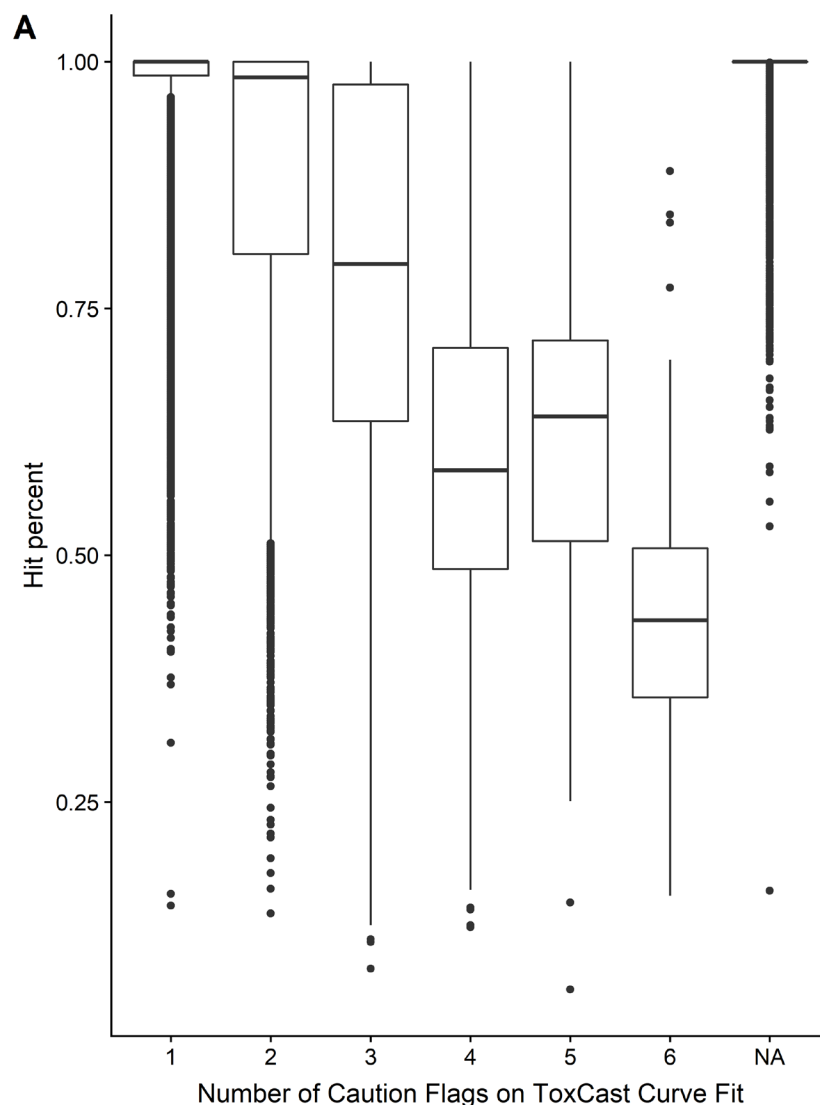
- ToxCast & Tox21 Summary Files for invitroDBv3.3:** Data for a single chemical endpoint pair for thousands of chemicals and assay endpoints for 20 variables such as the activity or hit call, activity concentrations, whether the chemical was tested in a specific assay, etc.
 - [Download ToxCast Summary Information](#)
 - [Download ReadMe](#)
- ToxCast & Tox21 Data Spreadsheet for invitroDBv3.3:** A spreadsheet of EPA's analysis of the chemicals screened through ToxCast and the Tox21 collaboration which includes EPA's activity calls from the screening of over 1,800 chemicals.
 - [Download Data](#)
 - [Download ReadMe](#)
- ToxCast Data Pipeline R Package (tcpl released with invitroDBv3.3):** The R computer programming package used to process and model all EPA ToxCast and Tox21 chemical screening data. The files include the R programming package as well as documents that provide overviews of the data analysis pipeline used and the R package. Users will need experience with R to use these files.
 - [Download Package](#)
 - [TCPL Overview](#)
- ToxCast Database (invitroDBv3.3):** Database of EPA's analysis of chemicals screened through ToxCast assays, includes a MySQL database and the R package used to process the data.
 - [Download Database](#)
 - [Download ReadMe](#)

Loading concentration response data from mc5 and mc6 from the database

```
chems <- tcplLoadChem()
mc5 <- tcplPrep0tpt(tcplLoadData(val=chems$spid, lvl=5, type = 'mc', fld='spid'))
mc6 <- tcplPrep0tpt(tcplLoadData(lvl=6, fld='m4id', val=mc5$m4id, type='mc'))
setDT(mc6)
mc6_mthds <- mc6[, .( mc6_mthd_id = paste(mc6_mthd_id, collapse=",")), by = m4id]
mc6_flags <- mc6[, .( flag = paste(flag, collapse=";")), by = m4id]
mc5$mc6_flags <- mc6_mthds$mc6_mthd_id[match(mc5$m4id, mc6_mthds$m4id)]
mc5[, flag.length := ifelse(!is.na(mc6_flags), count.fields(textConnection(mc6_flags), sep = ','), NA)]
```

- Mc5 and mc6 also available as .Rdata and .xlsx in latest release
- Questions to consider
 - Does the question require uniqueness by sample id (spid) or uniqueness by chemical? [see tcplSubsetChid() for uniqueness by chemical id]
 - How should sensitivity and specificity be balanced? I.e., how to permissively or stringently filter the data?

Filtering data for curve quality



- Using version 3.0 of invitrodb, curve-fits with 3+ caution flags seemed to demonstrate an obvious reduction in the reproducibility (or quality) of the curve-fit.
- Hit-percent is a summary metric indicating the % of bootstrap resampled curve-fits that were positive; derived from smooth nonparametric bootstrap resampling of the concentration-response data followed by curve-fitting (1000 resamples) [Watt & Judson, 2018, *Uncertainty quantification in ToxCast high throughput screening*. DOI: <https://doi.org/10.1371/journal.pone.0196963>].

Figure from Supplemental Appendix, Paul Friedman et al. (2019) Examining the utility of *in vitro* bioactivity as a protective point of departure: a case study. *Toxicological Sciences*. DOI: [10.1093/toxsci/kfz201](https://doi.org/10.1093/toxsci/kfz201)

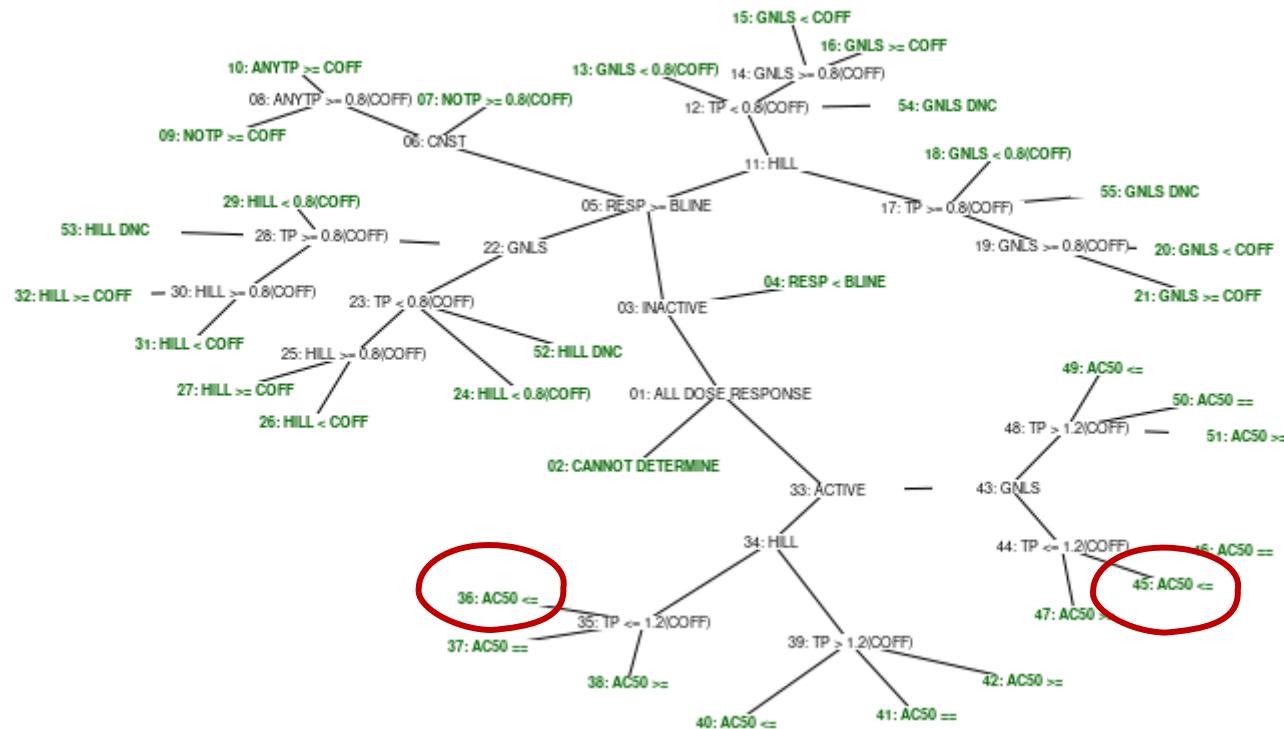
Filtering data for curve quality

- Could also consider using combinations of flags for fit-for-purpose dataset cleaning

```
> tcplMthdList(6)
  mc6_mthd_id      mc6_mthd                                     desc nldr
1:          6 singlept.hit.high      Look for single point hits with activity only at the highest conc tested  0
2:          7 singlept.hit.mid       Look for single point hits with activity not at highest conc tested  0
3:          8 multipoint.neg         Look for inactives with multiple medians above baseline  0
4:         10      noise              Look for noisy curves, relative to the assay  0
5:         11    border.hit           Look for actives with borderline activity  0
6:         12    border.miss          Look for inactives with borderline activity  0
7:         18 modlga.lowconc          AC50 less than lowest concentration tested  0
8:         15    gnls.lowconc         Look for low concentration gnls winners  0
9:         16    overfit.hit  Flag hit-calls that would get changed after doing the small N correction to the aic values.  0
10:        17    efficacy.50  Flag hit-calls with efficacy values less than 50% -- intended for biochemical assays.  0
11:        19    viability.gnls      Flag hit-calls with cell viability assay that are fit with gnls winning model  0
```

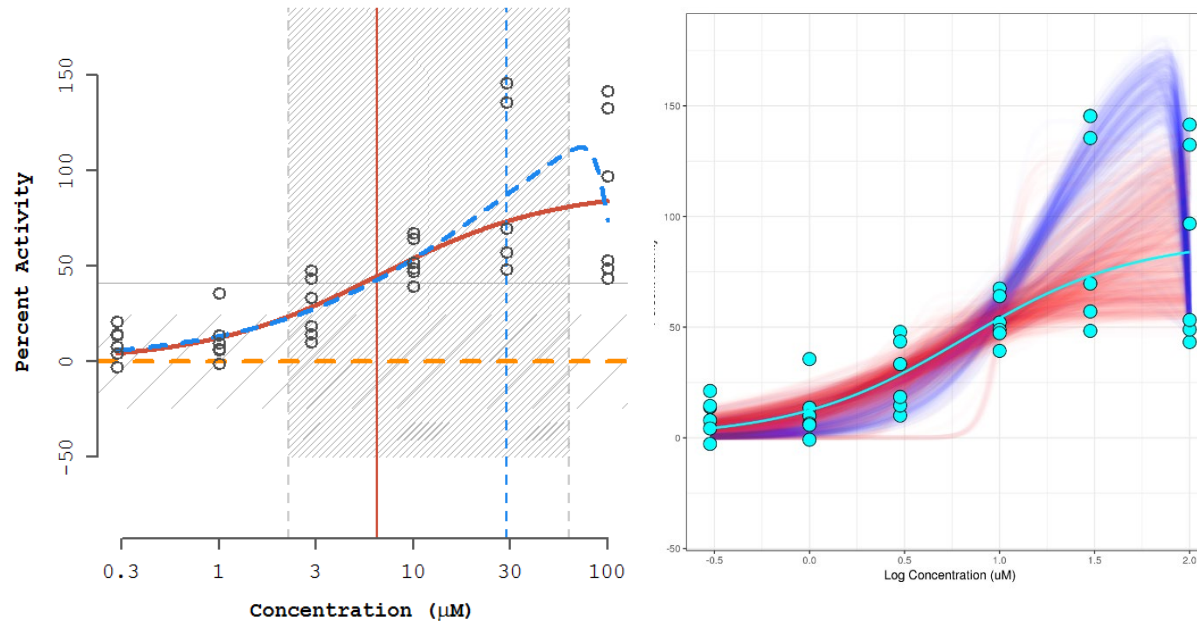
Filtering for the shape of the curve

Output from `tcplPlotFitc()`



- Could drop `fitc %in% c(36,45);` represent curve fits with very low efficacy and ($\leq 1.2 * \text{coff}$) and potency lower than the conc range screened ($\text{AC50} \leq \text{conc range screened}$).
- Can also use the maximum median response, concentration index, and the modeled potency to eliminate these types of low efficacy curve-fits that suggest potency in a range with no information about slope.

Using context from uncertainty quantification



- Using toxboot to resample datapoints from the curve for an m4id, with added noise (0 mean).
- Tcpl level 4 (mc4) fitting of resampled data.
- Repeat x1000.
- Store the information from each resampled fit.
- [Watt & Judson, 2018, *Uncertainty quantification in ToxCast high throughput screening*. DOI: <https://doi.org/10.1371/journal.pone.0196963>].

Mc7 field	Short Definition
M4id	Mc4 curve fit id
M7id	Mc7 toxboot id
Aeid	Assay endpoint id
Hit_pct	% of 1000 bootstrap resampled curve fits that were hitc==1
Modl_ga_min	Min log10-ac50
Modl_ga_max	Max log10-ac50
Modl_ga_med	Median log10-ac50
Modl_gw_med	Median log10 gain slope
Modl_ga_delta	The width of the modl_ga prediction (max-min)
Cnst_pct	% of 1000 fits that are constant
Hill_pct	% of 1000 fits that are Hill
Gnls_pct	% of 1000 fits that are Gain-loss

Filtering for cytotoxicity

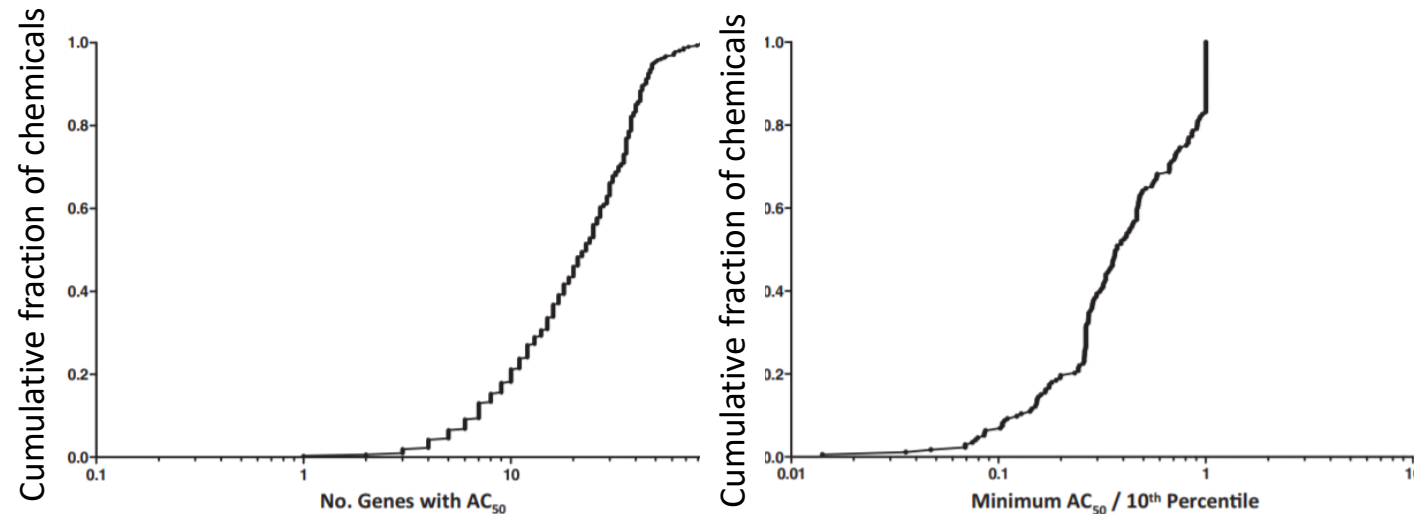
- Many approaches are possible.
- What is the stringency needed for the data application?
- Use parallel viability only, if available?
- Use a lower bound estimate of cytotoxic concentration?
 - How much lower must a bioactivity potency value be to suggest selectivity?
- The default `tcplCytoPt()` function is used to create the `invitrodb.cytotox` table (which is also released in `INVITRODB_V3_3_SUMMARY` as `CytoPt.xlsx` and `CytoPt.Rdata`).
 - Can customize `tcplCytoPt()` to use different assays to create a lower bound and median estimate.

Many of the substances in ToxCast appear non-selective

TOXICOLOGICAL SCIENCES **136**(1), 4–18 2013
doi:10.1093/toxsci/kft178
Advance Access publication August 19, 2013

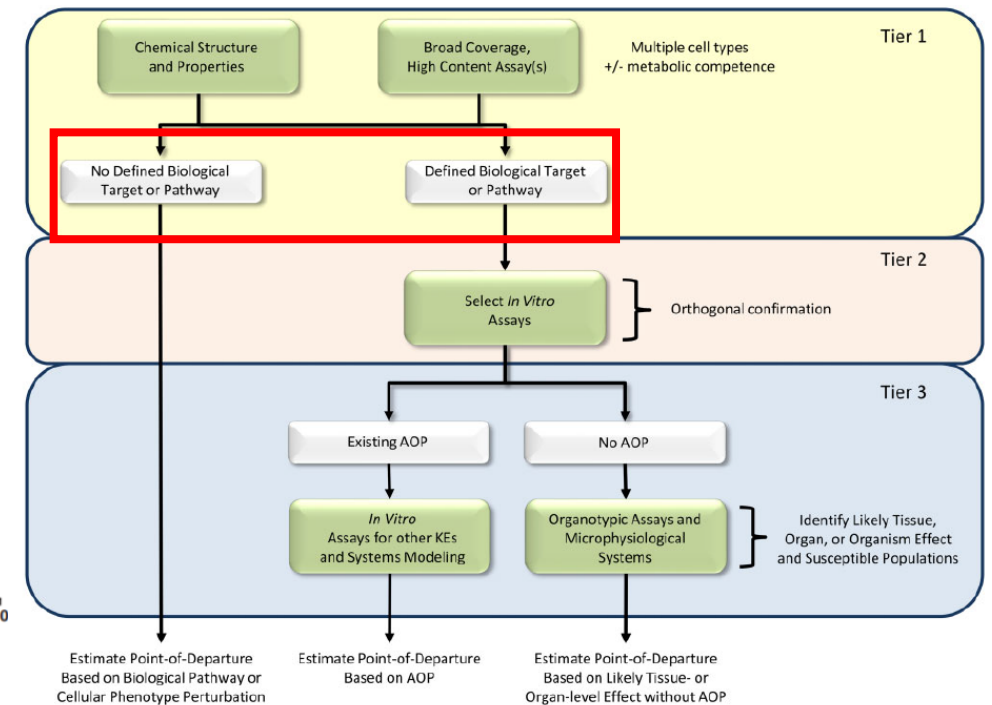
Incorporating New Technologies Into Toxicity Testing and Risk Assessment: Moving From 21st Century Vision to a Data-Driven Framework

Russell S. Thomas,^{*,†} Martin A. Philbert,[†] Scott S. Auerbach,[‡] Barbara A. Wetmore,^{*} Michael J. Devito,[‡] Ila Cote,[§] J. Craig Rowlands,[¶] Maurice P. Whelan,^{||} Sean M. Hays,^{|||} Melvin E. Andersen,^{*} M. E. (Bette) Meek,^{||||} Lawrence W. Reiter,[#] Jason C. Lambert,^{**} Harvey J. Clewell III,^{*} Martin L. Stephens,^{††} Q. Jay Zhao,^{**} Scott C. Wesselkamper,^{**} Lynn Flowers,[§] Edward W. Carney,[¶] Timothy P. Pastoor,^{‡‡} Dan D. Petersen,^{**} Carole L. Yauk,^{§§} and Andy Nong^{§§}

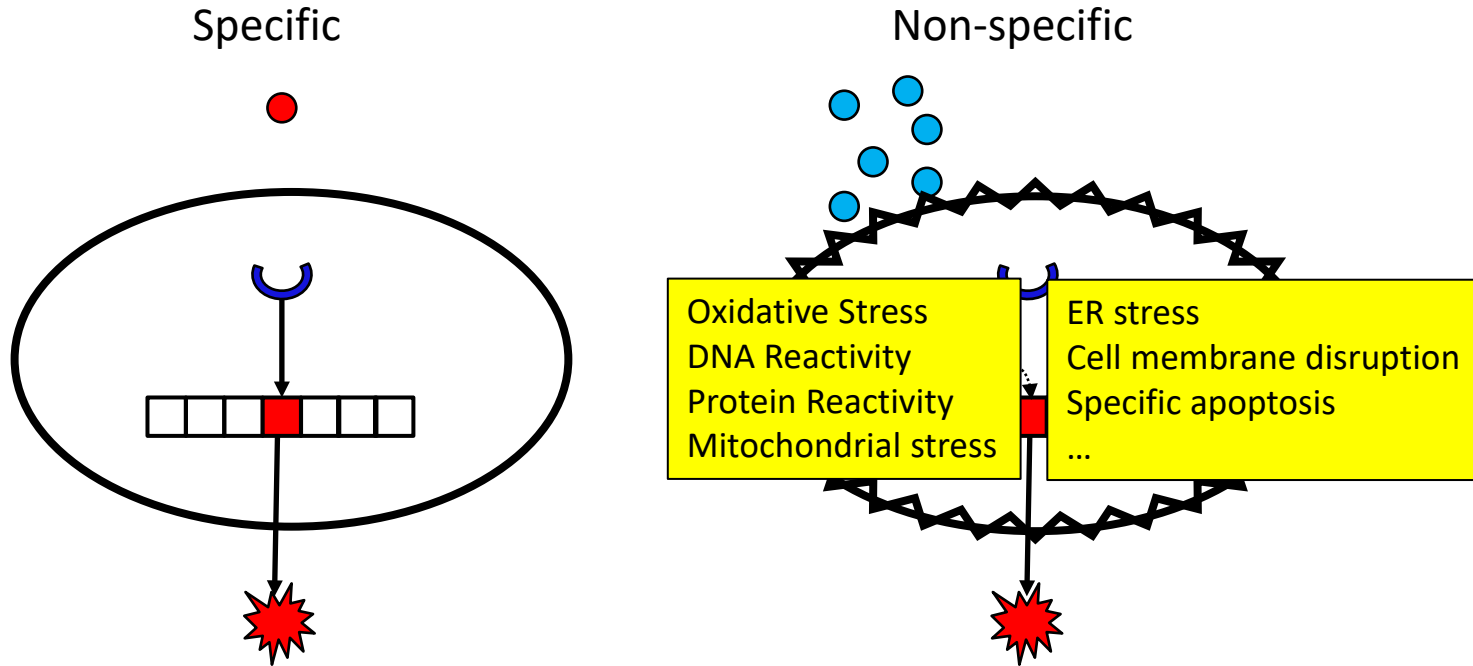


- Many chemicals appear to act at many targets, or be non-selective
- This could be used to subset chemicals into screening tracks

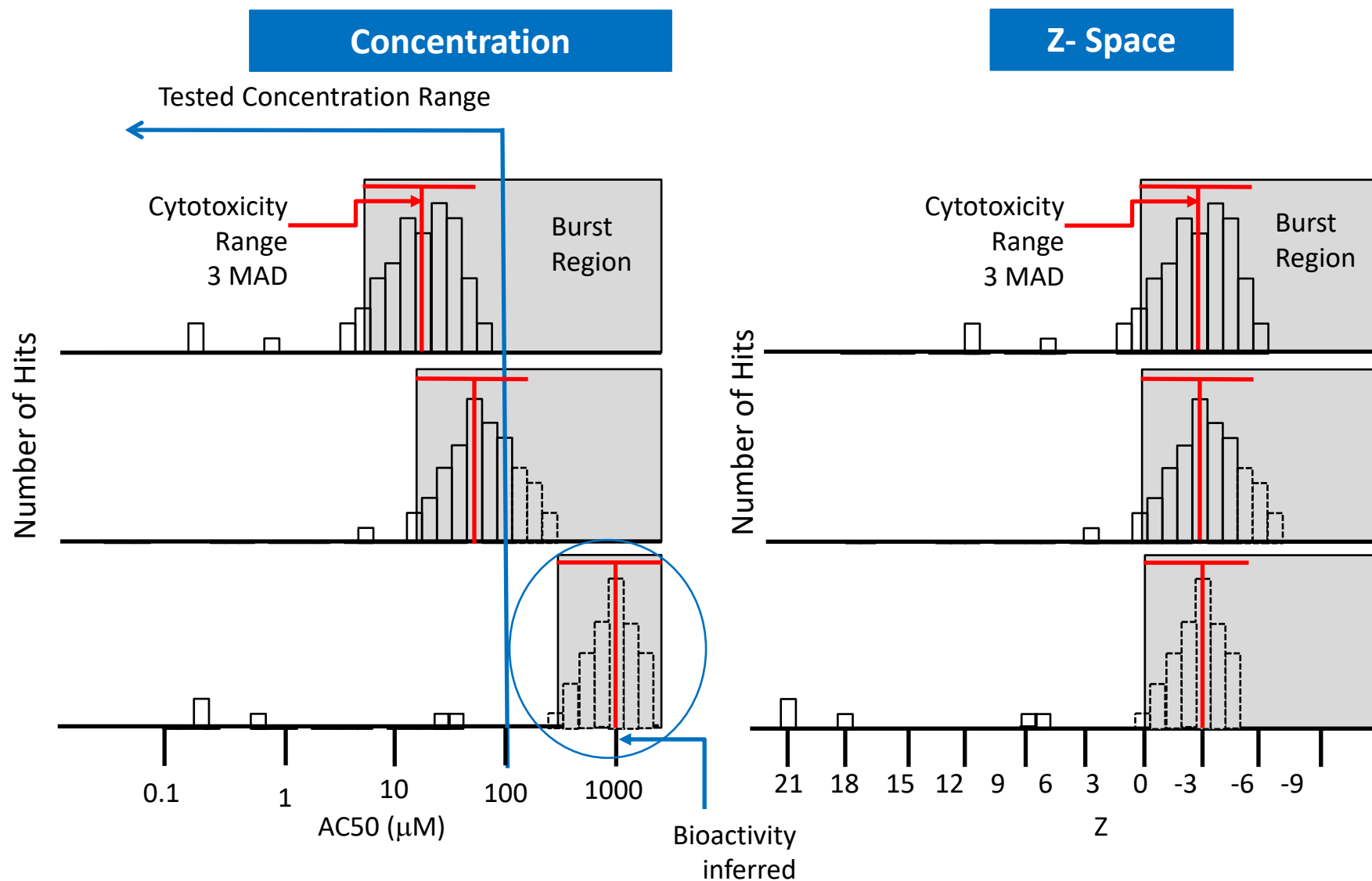
THOMAS ET AL. | 323



Schematic explanation of the burst



Most chemicals display a “burst” of potentially non-selective bioactivity near cytotoxicity concentration



The cytotoxicity “burst” is useful for context: the fine print

- The latest Comptox Chemicals Dashboard release (version 3.5, July 2020 release) demonstrates a cytotoxicity threshold based on the latest ToxCast database (invitrodb version 3.3, released Aug 2020). This value can change as more cytotoxicity data become available, curve-fitting approaches for existing data change, or the “burst” calculation approach is updated.
- In invitrodb version 3.3, 88 assays are considered for the cytotoxicity threshold. A positive hit must be observed in 5% of these assays (noting that not all chemicals are screened in all 88 assays) in order to assign a cytotoxicity threshold. The cytotoxicity threshold is a median of AC50 potency values from the N assays with a hit. The cytotoxicity threshold visualized in the Dashboard is a lower bound on this estimate, calculated as the median cytotoxicity potency minus 3 times the global median absolute deviation.
- This is discussed further in a publication ([10.1093/toxsci/kfw148](https://doi.org/10.1093/toxsci/kfw148)) and the ToxCast Pipeline R package (tcpl) function, tcplCytoPt() (available on CRAN: <https://cran.r-project.org/web/packages/tcpl/index.html>).
- **If fewer than 5 cytotoxicity assays demonstrate a positive hit, a default of 1000 micromolar is assigned for the chemical.**
- The **lower bound estimate of the cytotoxicity threshold or “burst” is useful context** for ToxCast results. Bioactivity observed below the cytotoxicity threshold may represent more specific activity that is less likely to be confounded by cytotoxicity.
- It is possible that AC50 values above the cytotoxicity threshold are informative. If an assay has a parallel cytotoxicity assay in the same cell type, that may be more informative for interpreting that assay. Or, if a result is consistent with an AOP relevant to the chemical with assay AC50 values above and below the cytotoxicity threshold, those data may be meaningful.

**Tcpl v3 is coming: Updates to enable
more curve-fitting updates**

Tcpl v3 will include a dependency of tcplFit2

- TcplFit2 (Judson and colleagues) functionality is based on BMDExpress
- Tcpl curve-fitting models (constant, Hill, and gain-loss models) are expanded to include Polynomial 1 (Linear), Polynomial 2 (Quadratic), Power, Exponential 2, Exponential 3, Exponential 4, and Exponential 5.
- Inclusion of these models impacts invitrodb in two primary ways:
 - (1) the need for long-format storage of generic modeling parameters and
 - (2) the need for updated logic on selection of winning model.
- Continuous hit call probability in tcplfit2 is calculated as the product of three different probabilities: median response and top of model exceed the cutoff, and AIC is less than the constant model.
- Plotting in tcpl v3 is expanded with a new utility called tcplPlot that provides interactive display of concentration-response curves through plotly integration with built-in REST api functionality. Integrating tcplPlot with tcpl allows for a consistent visualization of curves in reports and web applications.

Pivot invitrodb to accommodate tcplFit2

- Mc4 and mc5: added “long” format tables to accommodate many more outputs (current invitrodb schema is wide at mc4 and mc5 to accommodate only constant, gnls, and Hill)
- New model parameters from tcplFit2:
 - top_over_cutoff, rmse, a, er, bmr, bmdl, bmdu, caikwt, mll, hitcall, ac50, top, ac5, ac10, ac20, acc, ac1sd, bmd, tp, ga, b, p

Conclusions

- Tcpl and invitrodb provide a standard for consistent and reproducible curve-fitting and data management for diverse *in vitro* assay data with readily available documentation, thus enabling the sharing and use of these data in myriad toxicology applications.
- Continual improvements to the data (and the database) and the tcpl package itself.
- Now releases include all data as .Rdata and .xlsx such that the release files resemble the database.
- Datasets may need to be “cleaned” or refined on a fit-for-purpose basis to answer computational toxicology questions.
- The CompTox Chemicals Dashboard provides a data viewer, whereas tcpl and invitrodb provide tools for data science.

Acknowledgements

A huge team of people have contributed to the development, screening, and analysis of ToxCast and Tox21 data!

Key EPA ORD contributors to running ToxCast, tcpl, and invitrodb:

Jason Brown (brown.jason@epa.gov)

Madison Feshuk

Keith Houck

Richard Judson

Katie Paul Friedman (paul-friedman.katie@epa.gov)

Many, many assay data contributors

Collaborators

Menhang Xia and Ruili Huang and others at National Center for Advancing Translational Sciences



Center for Computational Toxicology and Exposure (CCTE)
Office of Research and Development (ORD)
US Environmental Protection Agency