2508/P867

Confidence in Fitting and Hitting Concentration-Response Data: Tox21 10k Library Pipeline Comparison

Sipes NS[1], Huang R[2], Shockley K[3], Martin MT[4], Shapiro A[1], Addington J[5], Auerbach SS[1], Paules R[1], Judson R[4], Houck K[4], Hong H[6], Hsieh JH[5]

[1]NTP/NIEHS/NIH, RTP, NC; [2]NCATS/NIH, Bethesda, MD; [3]NIEHS/NIH, RTP, NC;
[4]NCCT/USEPA, RTP, NC; [5]Kelly Government Solutions, RTP, NC; 6NCTR/FDA, Jefferson, AR

Abstract #2508

The Tox21 program has generated high-throughput screening data on thousands of chemicals. While the data are publicly available through partner websites, PubChem, and publications, the analyses are different. We developed a pipeline consensus to identify higher confidence chemical-assay calls and are developing a public web application. Tox21 chemical-assay pair activity calls (active, inactive and, in some cases, inconclusive) were compared among the 4 hit-call methods; CurveP and 3Stage from NIEHS, TCPL from the US EPA, and CurveClass from NCATS. Out of the 664,463 chemical-assay pairs (8,948 chemicals, 67 assays), 82% had total agreement (97% inactive), 13% had 3 pipelines in agreement where the agreement call was inactive (50%), active (23%), and inconclusive (27%), 4% had a 50/50 split, and the rest at 1%. High agreement assays were nuclear receptor agonist assays (e.g., androgen, thyroid, estrogen). Complete curves with high efficacies were overly represented in these assay curve fits. Antagonist assays were over represented in the assays with the lowest total agreement, this was expected since not all pipelines use the viability counter screen to adjust the calls. In addition, the lowest total agreement was associated with higher discordance among significance thresholds. Chemical purity did not appear to be an influential factor. Chemicals with the highest positive agreement included metals (e.g., phenylmercuric acetate, zinc pyrithione, and tributyltin chloride) occurring between 70-100 times, while chemicals with the lowest agreement (e.g., cycloheximide, fulvestrant and triclocarban) occurred about 15 times. It is important to note that some chemicals are more concordant in specific assays than others. In addition, median differences between calculated maximum efficacies and the pipeline's predetermined minimal activity significance threshold were lower when calls were discordant versus when all four pipelines agreed on an active call. Our comparative analysis provides the scientific community inclusive access and evaluation of Tox21 data with the ability to identify higher confidence activity calls across pipelines. *This abstract does not represent any US government policy.*

Background

Tox21 Federal Partnership

- Generating and compiling data toward better understanding and predicting toxicity
- Tested >8,000 chemicals in >60 high-throughput screening (HTS) assays at NIH/NCATS
- Multiple chemical-assay curve-fitting methods

Four concentration-response pipelines for Tox21 data

Name of Method	Institute	Public Access
3Stage[1]	NIH/NIEHS	method in PubMed
CurveClass[2-4]	NIH/NCATS	www.ncats.nih.gov & PubChem
CurvepwAUC[5]	NIH/NIEHS/DNTP	www.ntp.nih.gov
TCPL[6]	US EPA	www.epa.gov

- How do the four pipelines differ?
- What is the concordance among active 'hits' and inactive chemical-assay pairs?
- What parameters lead to greatest/least concordance?
- Can we derive more-confident calls using the four pipelines?

Goals

1. **Determine concordance among pipelines**
2. **Identify parameters leading to greatest/least concordance**
3. **Develop public web application to access all data**

Methods

1. Download data, and process for comparison (e.g., match assay names and readout columns).
2. Determine a simplified activity hit-call for each chemical-assay pair for each pipeline (HC) to allow comparisons across pipelines

simplified readout	3Stage	CurveClass	CurvepwAUC	TCPL
	FINAL_CALL	Activity	hitcall	hitc
active	ACTIVE_UP *(ACTIVE_DOWN)	active agonist *(active antagonist)	1	1
			0.5	
inactive	INACTIVE	inactive	0	0
	ACTIVE_DOWN *(ACTIVE_UP)	active antagonist *(active agonist)	-1	
			-0.5	
inconclusive	INCONCLUSIVE_UP	inconclusive	2	NA
	INCONCLUSIVE_ DOWN	inconclusive agonist	blank	
	TOTALLY_ INCONCLUSIVE	inconclusive antagonist		

*for assays measuring an antagonist readout

3. Calculate concordance for analyses and website
   1. Degree of agreement
      - Total (active and inactive) agreement, three agree, fifty-fifty , other
      - Simplified where inconclusive=inactive. Total (1111) and three (1110) active agreement, total (0000) and three (0001) inactive agreement, and fifty-fifty (0011).
   2. Calculate one consensus call (with confidence score) for each chemical-assay pair from each pipeline's simplified readout
      - **Consensus call (HC_M)**: Majority vote, with ties leaning toward activity (with at least two pipelines indicating active) or inconclusive (with at least two pipelines indicating inconclusive)
      - **Combined score**: $1-\frac{\sum_i^N |HC_i - HCM| \times n_i}{N}$, where N = number of unique $HC \neq HC_M$, n_i = number of times the HC_i appears over the four pipelines.
   4. Perform data analysis in R and Partek
   5. Develop staging website

Basic evaluation of approaches

	pipeline parameters		3Stage	CurveClass	CurvepwAUC	TCPL
per treatment (i.e., well-level)	data level		well-level	well-level	well-level	pooled well-level
	data fitting model		constant or Hill model	Hill model	model free	constant, gain-loss, or Hill
	automated pipeline		NO	YES	YES	YES
	outlier removal		YES (from CurveClass)	YES	YES	N/A
	response threshold		same as CurveClass	3SD of normalized responses in DMSO plates	Threshold (THR) to reduce POD variance, using 5-45%. Cutoff of 0.02 $\log_{10}$ of pooled SD of POD between two THRs is suggested	10 times the BMAD (Baseline Median Absolute Deviation) =MAD for first two concentrations across entire assay
	activity outcome per well	group	ACTIVE*[1]/ACTIVE*[2]/INCONCLUSIVE*[3]/INCONCLUSIVE*[-3]/ACTIVE*[-1]/ACTIVE*[-2]	curve class (+/- 1.1, 1.2, 1.3, 1.4, 2.1, 2.2, 2.3, 2.4; -3, 4, 5)	N/A	N/A
		activity metrics	AC50, efficacy	AC50, efficacy,CurveRank -9to9	AC50, POD, wAUC, efficacy	N/A
		flag	N/A	carry over	N/A	N/A
per source (Tox21 ID)	data collapsing method		majority vote; mean or median of Hill eq parameter estimates	average score based on Curve Rank + reproducibility groups	median value of the activity metrics	N/A
	known artifacts used to adjust data	carry over	No	Yes, assigning them as a Curve Class (=5) and a CO flag	Yes, correcting them to regular responses	N/A
		flare	No	Yes, by a pattern detection algorithm	No	No, but flags applied for row, column, pintool, chemical plate effects
		autofluor (bla)	No	Yes, by promiscuous activity in ch2 and autofluor data	Yes, by active/non-active info in ratio, ch2, autofluor data	No
		ratio/ch2 conflict	No	Yes, by activity in both ratio and ch2	Yes, by active/non-active info ratio, ch2, and ch1 data	No
		cytotox	No	Yes, by AC50 fold change difference + t-test in ratio and via data	Yes, by wAUC fold change difference in ratio and via data	Not incorporated, but a chemical-wise cytotox limit is calculated using the median AC50 w/ ≥2 actives across the ~35 cytotox assays in Tox21/ToxCast
	activity outcome per source (i.e., Tox21 ID)	group	**aggregated** / **formula for aggregation**	**aggregated** / **formula**: **Curve rank** / **Reproducibility**	**aggregated** / **formula**	**aggregated** / **formula**
			ACTIVE_UP / ACTIVE*[2] or ACTIVE*[1]	active agonist: >=1 / active match; >4 / inconclusive	active(1)	active (1) / wAUC > T_{50} & ≠ 0 (TSO: 50% percentile of wAUC distribution with EC50 > 10 µM) / when curves are > threshold and hill or gnls model wins
			ACTIVE_DOWN / ACTIVE*[-2] or ACTIVE*[-1]	active antagonist: <=-1 / active match; <-4 / inconclusive	active in irrelevant direction (-1)	N/A / wAUC < T_{50}*-1 & ≠ 0 & wAUC < 0 / N/A
			INACTIVE / INACTIVE*	inactive / >-1 and <1 / inactive match	inactive (0) / wAUC = 0	inactive (0) / when curves are < threshold or cnst model wins
			INCONCLUSIVE_UP / INCONCLUSIVE*[3]	inconclusive agonist: >=1 / mismatch; >=1 and <=4 / inconclusive	marginal active (0.5)	N/A / wAUC ≤ T_{50} & ≠ 0 / N/A
			INCONCLUSIVE_DOWN / INCONCLUSIVE*[-3]	Inconclusive antagonist: >=-4 and <=-1 / inconclusive; <=-1 / mismatch	marginal active in irrelevant direction (-0.5)	N/A / wAUC ≥ T_{50}*-1 & ≠ 0 & wAUC < 0 / N/A
			TOTALLY_INCONCLUSIVE / no majority	Inconclusive: >-1 and <1 / inconclusive	inconclusive (blank)	unable to fit (-1) / with flags related to artifacts and data quality / not enough unique concentrations to fit data
		potency	mean.log2EC50.nls	average AC50	AC50, POD (median value)	ACB, ACC, AC50
		efficacy	mean.RMAX.nls	average efficacy	Emax (median value)	Emax
		other	N/A	N/A	wAUC (median value)	N/A
		flag	N/A	Contamination, signal flare	a_normal/b_autofluor/c_contradict /d_cytotoxic/e_weaknoisy, one flag per source	up to 17 Flags with 8 req full conc response info (e.g, row/ column effects), multi-flag per source

CurvepwAUC is model-free
TCPL does not remove outliers
Response thresholds vary
3Stage uses majority vote
TCPL calculates per Tox21ID
CurveClass & CurvepwAUC calculate and applies flags to data
TCPL calculates some flags, but does not apply to data
TCPL does not make inconclusive calls
CurvepwAUC gives weighted Area Under Curve (wAUC)
Most provide goodness-of-fit and/or other issue flags

Table 1. Pipeline comparison. This table provides notable and differences among the four pipelines. Call-out boxes refer to some notable similarities and differences. 3Stage is a generalized concentration-response model that has been examined on performance considering homoscedastic and heteroscedastic error models, with a conservative estimate of activity when majority agrees. CurveClass is used for high-throughput screening data at NIH/NCATS and incorporates curve-class (shape of the curve) to separate out responses and adjusts outcomes based on artifacts. CurvepwAUC is a model-free method incorporating weighted area under the curve (wAUC), outliers and artifacts to make calls. TCPL is used in the US EPA's ToxCast program for fitting HTS data, using 3 models to fit at the source-level, thereby eliminating the need for a Bayesian outlier detection, referring the user to goodness-of-fit flags, allowing for user-interpretation.

Concordance among pipelines

663,737 Tox21ID - assay combinations (13,126 unique Tox21ID, 67 assays)

A Overall Calls **(in %)**

Call	Curve Class	3Stage	Curvep wAUC	TCPL	Combined call
inactive	83.1	85.0	87.4	90.3	86.6
inconclusive	11.0	11.6	5.9	n/a	6.2
active	5.9	3.4	6.7	9.7	7.2

B Overall Calls **(in #)**

Call	Curve Class	3Stage	Curvep wAUC	TCPL	Combined call
inactive	551,575	563,960	580,295	599,617	575,087
inconclusive	73,300	76,968	38,795	n/a	41,041
active	38,862	22,809	44,647	64,120	47,609

Table 2. Concentration-response activity calls in A) percent of total calls within a pipeline and B) total number of calls. A majority of calls are inactive across pipelines. Inconclusives make up 1.9xs, 3.4xs, and 0.9xs the actives (CurveClass, 3Stage, CurvepwAUC, respectively). TCPL pipeline has highest number of actives, while 3Stage has the lowest number of actives

3Stage	Curve Class	Curvep wAUC	TCPL	Counts
0	0	0	0	6E+05
0	0	0	1	19414
0	1	1	1	17096
1	1	1	1	15151
0	0	1	1	6372
0	1	0	1	2516
1	0	0	0	2348
0	0	1	0	2316
1	0	0	1	1601
1	0	1	1	1262
0	1	1	0	1164
0	1	0	0	980
1	1	1	0	794
1	1	0	1	708
1	0	1	0	492
1	1	0	0	453

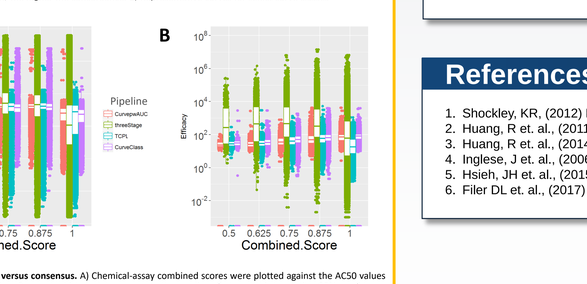
0000 - Top scenario is when all agree with inactive call

0001 - 84% are due to the 3 inactive pipelines being inconclusive
- 86% had combination of TCPL flags: <50% efficacy, borderline active, only one concentration above baseline) or no flag
- In case of no flag, 99% of times 3 other pipelines were all inconclusive, 87% were antagonist assays, 83% of time cytotoxicity was noted (CurvepwAUC)

0111 - >99% is when 3Stage is inconclusive

Table 3. Count of each permutations of pipeline activities when inconclusives are assumed inactive

Global Concordance: 545,178 chemical-assay pairs (82%) where all four pipelines agree. Of these, 530,027 (97%) are inactive, 15,151 (3%) are active.



0000	0001	0011	0111	1111	not tested
591,070	25,058	12,598	19,860	15,151	215,705

Assay Type: viability, antagonist, agonist

Figure 1. Concordance across assays and chemicals. Inconclusive calls were turned into inactive calls and concordance was determined. Five scenarios are shown 1) all four pipelines had active agreement (purple), three actives (blue), fifty-fifty (yellow), three inactives (red), and all inactive agreement (gray).

Discussion

The analysis performed represents a way to provide confidence in activity calls based on four separate and independent pipelines for fitting concentration-response curves to the Tox21 data. We have identified:

- Top agreement scenarios
  1. When all four pipelines agree (i.e., 0000)
  2. When TCPL calls the chemical-assay pair active and the rest call it inactive.
     - Mostly due to other pipelines calling these inconclusive due to calling cytotoxicity a factor in the antagonist assay
  3. Third is when 3Stage is inactive (>99%) inconclusive and others are active, indicating conservatively calling actives
- Antagonist assays have lower overall agreement (due to cytotoxicity call adjustment)
- High standard deviation across the four pipeline's cutoff value tend to give lower total agreement for antagonist and viability assays.
- Chemical purity, efficacy, or AC50 does not correlate with pipeline concordance
- Shape of the curve (via curve class) correlates with pipeline concordance

Take home:
- Pipelines in general agree.
- When reviewing HTS data, evaluate concentration-response curves.

Parameter evaluations

Total call agreement over assays range from 64% - 95%

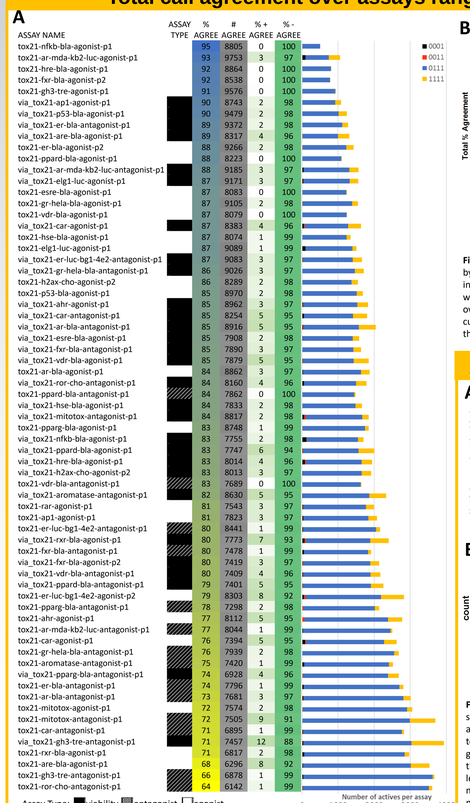


Figure 2. Assay agreement A) over individual assays categorized by % total agreement, with values, % active (1111) and inactive(0000) agreement, and plotting of the four scenarios with at least one active call (i.e., 0001, 0011, 0111, 1111). B) over assay type, with data colored by the average hit-calling cutoff among the four pipelines and data point size according to the standard deviation of the cutoff among the four pipelines.

Agreement over chemicals

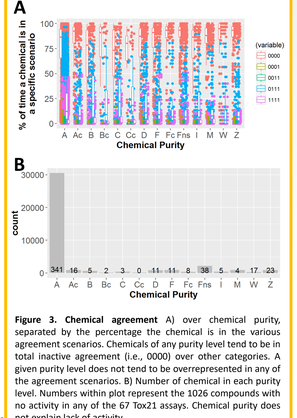
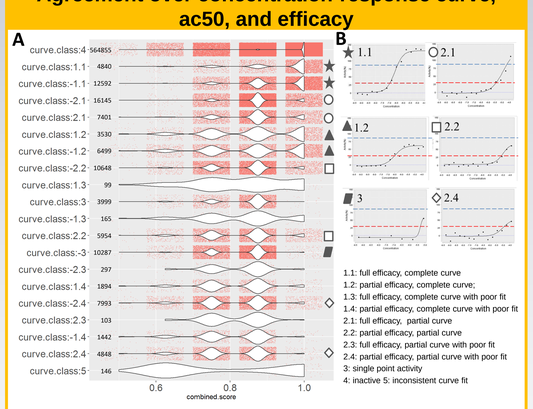


Figure 3. Chemical agreement A) over chemical purity, separated by the percentage the chemical is in the various agreement scenarios. Chemicals of any purity level tend to be in total inactive agreement (i.e., 0000) over other categories. A given purity level does not tend to be overrepresented in any of the agreement scenarios. B) Number of chemical in each purity level. Numbers within plot represent the 1026 compounds with no activity in any of the 67 Tox21 assays. Chemical purity does not explain lack of activity.

Agreement over concentration-response curve, ac50, and efficacy



1.1: full efficacy, complete curve
1.2: partial efficacy, complete curve;
1.3: full efficacy, complete curve with poor fit
1.4: partial efficacy, complete curve with poor fit
2.1: full efficacy, partial curve
2.2: partial efficacy, partial curve
2.3: full efficacy, partial curve with poor fit
2.4: partial efficacy, partial curve with poor fit
3: single point activity
4: inactive 5: inconsistent curve fit

Figure 4. Concentration-response shapes versus consensus. A) Chemical-assay combined scores (i.e., degree of consensus, with 1 representing all 4 pipelines agree (e.g., 0000 and 1111), and values <1 representing lower agreement across pipelines) were plotted over the different concentration-response curve shapes, represented by curve classes 1-5. Values on the left represent total number of chemical-assay pairs with the specified curve class. In general, the more complete curve with good fit, the higher the concordance. B) Representative curves for select curve classes

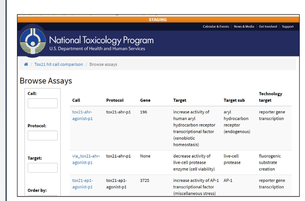
Figure 5. AC50 and efficacy versus consensus. A) Chemical-assay combined scores were plotted against the AC50 values calculated using each pipeline. Box and whisker plots represent the median, first and third quartile, 95% confidence interval, and outlier points. Median AC50s are generally consistent across the pipelines over consensus, with larger variance in the 3Stage. B) Chemical-assay combined scores were plotted against the maximum efficacy calculated using each pipeline. Values were generally consistent for the TCPL, CurveClass, and CurvepwAUC pipelines.

Public access and future efforts

The Tox21 Hit Call Comparison web application is intended to provide a publicly available web application to access and compare the Tox21 data across the four pipelines.

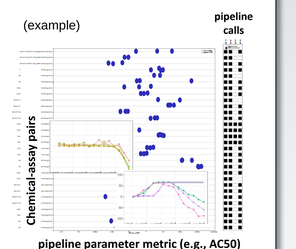
Information includes:

Browse or search by assays (including assay annotation information (e.g., target, technology, known chemical artifacts, cell line, positive control & concentration, incubation time) OR Browse by chemical



View comparisons by Tox21ID or chemical, with ability to evaluate curves

(example)



References

1. Shockley, KR, (2012) Environ Health Perspect 120, 1107–1115.
2. Huang, R et. al., (2011) Environ Health Perspect 119, 1142–1148.
3. Huang, R et. al., (2014) Sci Rep 4:5664, 1-9.
4. Inglese, J et. al., (2006) PNAS 103, 11473–11478.
5. Hsieh, JH et. al., (2015) J Biomol Screen 20, 887-97.
6. Filer DL et. al., (2017) Bioinformatics 15, 618-620.

Nisha S. Sipes | nisha.sipes@nih.gov | 984-287-3187