Correlation of Tox21 and ToxCast *In Vitro* and Small Model Organism Outcomes to Rat Oral Toxicity

W Polk[1], P Ceger[1], X Chang[1], N Kleinstreuer[1], J Strickland[1], M Paris[1], D Allen[1], W Casey[2]

[1]ILS/NICEATM, RTP, NC, USA; [2]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA

Abstract

At present, many national and international regulatory authorities use data from rat acute oral toxicity test methods for hazard classification and labeling. The Tox21 and ToxCast programs have tested over 8,000 and 1,800 chemicals, respectively, *in in vitro* and zebrafish (ZF) assays. We evaluated data from Tox21 and ToxCast to determine the potential of the more than 800 measures collected thus far to reduce animal use in toxicity testing for hazard identification. Rat oral LD_{50} data were obtained for 3,582 Tox21 and 1073 ToxCast Phase I and II chemicals. An ongoing analysis identified high-quality LD_{50} data for 76 chemicals that have been tested in ZF toxicity assays. The Tox21 and ToxCast data were analyzed for correlation and model fit to the LD_{50} data in order to determine which tests (and combinations thereof) best characterized the rat oral toxicity data. Correlation analyses were performed on binary outcomes of response for chemicals classified by LD_{50} as "toxic" (LD_{50} < 5000 mg/kg-bw). In this assessment of fit to the rat oral LD_{50} results, some models returned a sensitivity >0.46, which was modestly improved by including assays identified through random forest assessment. In parallel with the *in vitro* assessment, ZF toxicity assays were found to be more sensitive than rat oral toxicity for 75 of 76 chemicals, which was confirmed with a Mann–Whitney U test ($p < 1^{-15}$). Correlating the combined *in vitro* assays to rat oral LD_{50}s suggests that combinations of *in vitro* assays and small model organisms offer promise for predicting outcomes of rat acute LD_{50} limit tests. *(Data in poster abstract have been updated to reflect the most recent analyses.)*

Introduction

- Traditional acute oral toxicity tests yield an LD_{50} value, the dose of a test chemical that causes death in 50% of test animals during a 14-day observation period following a single, gavage-administered dose. LD_{50} data are used in a variety of regulatory applications for chemical hazards, including developing appropriate hazard labeling, product usage guidelines, personal protective equipment requirements, and transportation restrictions.
- There are thousands of chemicals in commerce that lack sufficient testing data.
- The Tox21 and ToxCast programs are working to address this problem using quantitative high-throughput screening (qHTS) assays to help understand how human biology is impacted by exposure to chemicals and to determine which exposures are the most likely to lead to adverse health effects.
- In this project, we compared data from several of the completed phases of these programs to rat oral LD_{50}s to determine whether these data could be used as an alternative to acute toxicity testing. Each dataset was analyzed by two methods:
  1. Correlation was calculated for the continuous variables.
  2. Correlation, sensitivity, and specificity were calculated on a binary transformation of the data as compared to the rodent oral LD_{50}.

Data Sources

NICEATM LD_{50} Database

- The National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) collected rat oral LD_{50} values for 3,884 unique chemicals from the following sources:
  1. NICEATM pesticide actives database (data obtained from the U.S. Environmental Protection Agency [EPA]) (n = 46)
  2. ChemID Plus (n = 3,299)
  3. European Chemicals Agency (n = 374)
  4. EPA Pesticide Reregistration Eligibility Decisions (n = 3)
  5. U.S. Hazardous Substances Databank (n = 162)
- All values identified were used in our analyses as they were reported.
- If a single source included multiple LD_{50} values for a single chemical, the lowest LD_{50} value was selected.

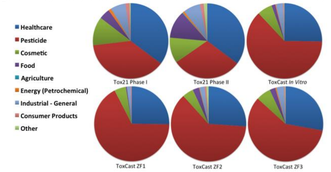
High-Throughput Data

- Tox21 is a U.S. federal interagency collaboration (Tice et al. 2013) in which qHTS methods are being used to evaluate the biological activity of >8,000 compounds and to map the observed activities to toxicity pathways. Two unique datasets from Tox21 were included in this analysis:
  1. **Tox21 Phase I** includes cytotoxicity assays using 11 cell types.
  2. **Tox21 Phase II** includes assays that cover over 30 cell signaling pathways.
- The EPA ToxCast program (Judson et al. 2010) has tested approximately 1,800 chemicals in over 700 assays. The Tox21 Phase II assays are included in ToxCast, but were analyzed separately for this poster.
- Four unique datasets from ToxCast were included in this analysis:
  1. **ToxCast *In Vitro* Dataset** includes >700 cell-free biochemical and human cell assay endpoints.
  2. **Embryonic Zebrafish (ZF) Dataset 1** includes toxicity and malformation assessments of ZF exposed to test chemicals across a concentration range (Padilla et al. 2012).
  3. **Embryonic ZF Dataset 2** includes toxicity and malformation assessments of dechorionated ZF exposed to chemicals across a concentration range (Truong et al. 2014).
  4. **Embryonic ZF Dataset 3** includes toxicity and malformation assessments of dechorionated ZF that were exposed to chemicals at a single concentration. Data were provided as the percentage of the embryos displaying an outcome (Truong et al. 2014).

Test Set Generation and Characterization

- The chemicals in the six qHTS datasets were cross-referenced with chemicals in the rat oral LD_{50} database to produce six test sets, unique in size (**Table 1**) and chemical space (**Figure 1**).
  - Regulatory categorization was applied to each chemical using ACToR and ChemID+ descriptors.
  - Where multiple categories existed, the descriptor representing the context in which an LD_{50} value is most likely to be applied was used.

Figure 1. Regulatory Category Distributions of the Chemicals in the Analyses



Abbreviation: ZF = zebrafish.

Table 1. Source Data Description

Data Source	Number of Tests	Total Chemicals Tested	Number of Chemicals in Source Data with LD_{50}
Tox21 Phase I	13	2800	796
Tox21 Phase II	43	8597	3293
ToxCast *In Vitro* Assays	776[a]	1877	1073
ToxCast Zebrafish Dataset 1	3[b]	310	114
ToxCast Zebrafish Dataset 2	18	1064	792
ToxCast Zebrafish Dataset 3	22	424	325

[a] The number of tests differs from the number of assays because some assays provided multiple endpoints. For example, the mitochondrial membrane potential assay produced two endpoints, which differ by directionality of the response from baseline.

[b] Outcomes were combined into three variables prior to collection by NICEATM. For full list of assessments and combination criteria, see Padilla et al. (2012).

Data Processing

- The LD_{50} and qHTS data were transformed for analysis as follows:
  - For assessment of continuous variables in the Tox21 *in vitro* datasets, each rodent LD_{50} and qHTS point of departure (POD) was inverted and then log transformed ($\log_{10}[1/x]$).
  - For assessment of continuous variables in the ToxCast *in vitro* dataset, we used log half-maximal effective concentration (AC_{50} in µM) and log LD_{50}.
  - Nontoxic responses in the *in vivo* assay (LD_{50} > 5000 mg/kg) and non-responses in the HTS assays were assigned values corresponding to doses or concentrations, respectively, beyond the test range.
  - For prediction of the limit test outcome, each LD_{50} was converted to a binary value that reported whether the value was higher than 5000 mg/kg. Each qHTS outcome was converted to a binary value that reported whether a POD was established for the dose range tested (any response).
- Pearson's correlation was used to calculate coefficients of correlation for the qHTS assay outcomes and the rat oral LD_{50}s for both continuous and limit tests.
  - Sensitivity and specificity were calculated for the continuous and limit tests to determine the performance of the alternative assays to classify a chemical as "toxic" ($LD_{50} \leq 5000$ mg/kg) using the equations below:

$$Sensitivity = \frac{\#\ of\ true\ positive}{\#\ of\ true\ positive + \#\ of\ false\ negative} \quad (1)$$

$$Specificity = \frac{\#\ of\ true\ negative}{\#\ of\ true\ negative + \#\ of\ false\ positive} \quad (2)$$

- Random forest (RF) modeling was used to rank the relative importance of the ToxCast assays in predicting acute systemic toxicity.
  - RF modeling is a machine-learning technique based on randomized decision trees. The outputs of all trees are aggregated to obtain one final prediction based on the outcome with the lowest prediction error.
  - To avoid using missing data, the RF analysis was restricted to 313 ToxCast assays that tested the highest number of chemicals (612 chemicals). RF was performed with 500 iterations.
- The Mann–Whitney U test, a nonparametric test to determine whether two groups are different, was performed on the rat oral and ZF data.

Performance of Individual *In Vitro* Assays

- For the qHTS datasets, the individual assays with the highest correlations to rat oral LD_{50}s had correlation coefficients ranging from 0.01 to 0.24 (**Table 2**). The continuous analyses produced higher correlation coefficients than the limit test analyses.
  - Sensitivity for the individual assays with the highest correlations ranged from 0.09 to 0.43.
  - Specificity for the individual assays with the highest correlations ranged from 0.86 to 0.95.

Table 2. Performance Metrics for Highest Correlated Tests from In Vitro Data Sources

Data Source	Assay Name	Assay Descriptor	Correlation Coefficient (Continuous)	Correlation Coefficient (5000 mg/kg Limit)	Sensitivity (5000 mg/kg Limit)	Specificity (5000 mg/kg Limit)
Tox21 Phase I	HEK293	Human kidney	0.24	0.01	0.09	0.95
Tox21 Phase II	ARant_ HEK293	Androgen receptor	0.18	0.02	0.15	0.86
ToxCast *In Vitro*	BSK_4H_Pselectin_down assay	P_selectin	0.21	0.15	0.43	0.87

Assessment of Combined *In Vitro* Assays

- The continuous data from the Tox21 Phase I, Tox21 Phase II, and ToxCast *in vitro* assays were ranked by correlation to rat oral LD_{50}s. The six assays from each source with the highest correlations are presented in **Table 3**.

Table 3. In Vitro Assays with Highest Correlation to Rat Oral LD_{50}

Correlation Rank	Tox21 Phase I	Correlation Coefficient (Continuous)	Tox21 Phase II	Correlation Coefficient (Continuous)	ToxCast *In Vitro*	Correlation Coefficient (Continuous)
1	HEK293	0.24	ARant_ HEK293	0.18	BSK_4H_ Pselectin_down assay	0.21
2	BJ	0.24	p53_ HCT116	0.17	BSK_3C_ Eselectin_down	0.19
3	N2a	0.23	TRant_ GH3	0.17	BSK_hDFCGF_ Proliferation_down	0.18
4	Jurkat	0.22	ARE_ HEPG2	0.16	BSK_hDFCGF_ VCAM1_down	0.18
5	SKN-SH	0.22	AHR_ HEPG2	0.15	BSK_LPS_ CD40_down	0.18
6	H4iie	0.22	PPARgant_ HEK293	0.15	BSK_SAg_ Eselectin_down	0.17

- The six highest performing tests from each dataset were then combined into a single variable that reported the most sensitive outcome (lowest POD or AC_{50}). Performance was assessed for the combined variable against the rat oral LD_{50}s using both continuous variables and limit tests (**Table 4**).
  - Selection of the top six Tox21 tests by correlation coefficient increased sensitivity and decreased specificity compared with the best individual tests in **Table 2**.

Table 4. Performance Metrics for Combined Variables that Best Predict Rat Oral LD_{50}

Data Source	Number of Assays Used	Assay Identification Method	Correlation Coefficient (Continuous)	Correlation Coefficient (5000 mg/kg Limit)	Sensitivity (5000 mg/kg Limit)	Specificity (5000 mg/kg Limit)
Tox21 Phase I	6	Correlation	0.25	0.04	0.21	0.83
Tox21 Phase II	6	Correlation	0.22	0.12	0.26	0.85
ToxCast *In Vitro*[a,b]	6	Correlation	0.19	0.14	0.50	0.66

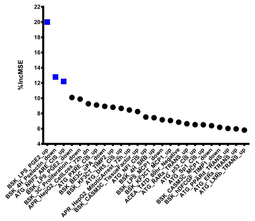
[a] The six top performers (in order) based on continuous analysis were BSK_4H_Pselectin_down, BSK_3C_Eselectin_down, BSK_hDFCGF_Proliferation_down, BSK_hDFCGF_VCAM1_down, BSK_LPS_CD40_down, and BSK_SAg_Eselectin_down.

[b] The six top performers (in order) based on limit analysis were BSK_hDFCGF_Proliferation_down, BSK_4H_Pselectin_down, BSK_hDFCGF_VCAM1_down, BSK_3C_Eselectin_down, BSK_hDFCGF_IP10_down, and BSK_3C_Vis_down.

Assessment of Combined *In Vitro* Assays (cont'd)

- Additional methods were applied to the ToxCast *in vitro* dataset to identify the assays with the best performance because this dataset had the highest number of *in vitro* tests. **Figure 2** shows the 25 most important ToxCast assays for predicting acute toxicity from the RF analysis.
  - The top three assays were selected as the top performing assays for later analyses.

Figure 2. ToxCast Tests Assessed by Random Forest Variable Importance



Abbreviations: %IncMSE = percent increase in mean squared error.

Blue squares identify the three assays that produced the highest percent increase in mean squared error when removed from the model.

- The continuous variables from the ToxCast *in vitro* datasets were optimized combining the top three tests identified by the RF analyses with the top six tests identified by the correlation analysis. The results were combined into a single variable that reported the lowest AC_{50} for each chemical (**Table 5**).
  - The top three ToxCast tests by RF ranking returned a correlation of 0.18, sensitivity of 0.46, and specificity of 0.63.
  - Combining the top six tests by correlation with the top three tests by RF analysis produced a total of eight assays because BSK_4H_Pselectin_down was included in both sets. The eight assays returned a correlation of 0.19, sensitivity of 0.51, and specificity of 0.61.

Table 5. Optimized ToxCast Prediction Performance

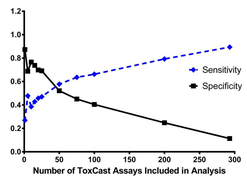
Data Source	Number of Assays Used	Assay Identification Method	Correlation Coefficient (Continuous)	Correlation Coefficient (5000 mg/kg Limit)	Sensitivity (5000 mg/kg Limit)	Specificity (5000 mg/kg Limit)
ToxCast *In Vitro*	3	RF	0.18	0.09	0.46	0.63
ToxCast *In Vitro*	8	Correlation and RF	0.19	0.10	0.51	0.61

Abbreviations: RF = random forest.

Optimization of Balanced Accuracy for the Continuous ToxCast *In Vitro* Data

- To determine the optimum number of ToxCast assays for comparison to LD_{50} data, the sensitivity and specificity of the highest performing (according to the continuous correlation coefficient) N assays was graphed for multiple Ns.
- The intersection point, which represents the best balance between sensitivity and specificity (balanced accuracy), occurred at the 45 tests with the highest performance. Sensitivity was 0.55 and specificity was 0.57 (**Figure 3**).

Figure 3. ToxCast Performance Assessed by Number of Included Tests



Performance of ToxCast Zebrafish Assays

Lethality

- ZF mortality by concentration response (Dataset 1 or 2) or by percent response of test animals (Dataset 3) resulted in correlation coefficients ranging from -0.02 to 0.14 and variable sensitivity (range of 0.10 to 0.43) and specificity (range of 0.59 to 0.92) for predicting rat oral LD_{50} values (**Table 6**).

Table 6. Performance Metrics for Lethality in Predicting Rat Oral LD_{50}s

Data Source	Correlation Coefficient (Continuous)	Sensitivity (5000 mg/kg Limit)	Specificity (5000 mg/kg Limit)
ZF Dataset 1	-0.02	0.43	0.66
ZF Dataset 2	0.04	0.42	0.59
ZF Dataset 3	0.14	0.10	0.92

Abbreviation: ZF = zebrafish.

All Endpoints

- The most sensitive ZF endpoint obtained by concentration response (Dataset 1 or 2) or by percent response of test animals (Dataset 3) was used for predicting rat oral LD_{50} is shown in **Table 7**.

Table 7. Performance Metrics for All Endpoints in Predicting Rat Oral LD_{50}s

Data Source	Correlation Coefficient	Sensitivity	Specificity
ZF Dataset 1	0.16	0.64	0.50
ZF Dataset 2	0.04	0.57	0.46
ZF Dataset 3	0.15	0.33	0.65

Abbreviation: ZF = zebrafish.

Post-Hoc Analysis

- Pairwise analysis of ZF toxicity with rat oral LD_{50}s demonstrated that:
  - When a ZF test was positive, the LC_{50} (mmol/L) was lower than the acute rat oral LD_{50} (mmol/kg) in 75 of the 76 true positives.
  - The lower LC_{50} response in ZF was confirmed to be significant with a Mann–Whitney U test (p < 1e-15).

Conclusions

- Alternative methods vary widely in their performance in predicting LD_{50} values.
- Our results indicate that increasing the number of endpoints by combining assay outcomes increases sensitivity, but at the expense of decreased specificity.
  - The number of tests and selection criteria used to identify tests impacts the performance of alternative test data for predicting *in vivo* acute toxicity.
  - Our data suggest an optimal number of between 6–45 assays for current datasets.
  - Use of multiple assays is consistent with current understanding of the relationship between individual endpoint assay outcomes and lethality. Individual endpoint assays measure a response of a single mechanism while lethality may occur as a result of a number of different mechanisms (cytotoxicity, inhibited blood clotting, neural transmission interruption, etc.).
- The individual endpoint assay responses seem to be predictive of the magnitude of the *in vivo* response, as demonstrated by the higher correlation obtained for predictions of the continuous variables as compared to those performed on the limit variables (**Tables 2** and **4**).
- The performance of these alternative assays cannot be compared between datasets because:
  - There are different numbers of chemicals included in each dataset.
  - There are different chemical categories included in each dataset.
    - Bias in chemical space coverage may impact the performance. For example, the ToxCast *in vitro* contained a large numbers of endocrine disruptors in that chemical library (EPA 2012).

Future Activities

- Work is currently underway to identify assays that improve the performance of prediction of highly toxic chemicals with specific molecular/physiologic targets, as these chemicals could be a primary reason for poor performance at the higher toxicity categories.
  - Neurotoxicity: The datasets are known to contain cholinesterase inhibitors, sodium channel modulators and agents that alter action potentials *in vivo*.
  - Cardiotoxicity: Cardiac glycosides have been identified in the datasets.
  - Vascular / blood toxicity: Agents that block clotting have been identified in the datasets.
- Quantitative structure–activity relationship modeling is being used to improve these predictions.

References

EPA. 2012. Endocrine Disruptor Screening Program Universe of Chemicals and General Validation Principles [Internet]. Washington, DC: U.S. Environmental Protection Agency.

Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, et al. 2010. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. Environ Health Perspect 118(4): 485-92.

Padilla S, Corum D, Padnos B, Hunter DL, Beam A, Houck KA, et al. 2012. Zebrafish developmental screening of the ToxCast™ Phase I chemical library. Reprod Toxicol 33(2): 174-87.

Tice RR, Austin CP, Kavlock RJ, Bucher JR. 2013. Improving the human hazard characterization of chemicals: a Tox21 update. Environ Health Perspect 121(7):756-65.

Truong L, Reif DM, St Mary L, Geier MC, Truong HD, Tanguay RL. 2014. Multidimensional in vivo hazard assessment using zebrafish. Toxicol Sci 137(1):212-33.

Acknowledgements

The Intramural Research Program of the National Institute of Environmental Health Sciences (NIEHS) supported this poster. Technical support was provided by ILS under NIEHS contract HHSN273201400003C.

The views expressed above do not necessarily represent the official positions of any Federal agency. Since the poster was written as part of the official duties of the authors, it can be freely copied.

A summary of NICEATM activities at the 2015 SOT Annual Meeting is available on the National Toxicology Program website at http://ntp.niehs.nih.gov/go/742110.