[TopCoder]

Marathon Match

Introduction to ToxCast Data v 1.0

3 April 2014

1.0 Introduction

This document aims at providing a simplified understanding of the data sets to be used in the LEL Prediction Challenge Marathon Match for ToxCast Project. As a lot of data is domain (biology and toxicity) specific, it is deemed fairly complex to comprehend all the information from the raw data and this guide is expected to help you gain necessary insights into the data that will help build successful models for the challenge.

The document starts with the organization of the data and provides an understanding of different data sets provided in various files and directories. Further, it provides some helpful information on the data sets that have been found to be important through initial analysis so to help get you started. Finally, it provides a fairly long list containing the description/explanation of various features/values that you would encounter in the data.

1.1 Project Overview

ToxCast is a project of the U.S. EPA whose goal is to develop methods to use in vitro assay chemical structure to predict the toxicity of environmental chemicals. This data consists of information on chemicals and in vitro assays.

[***In vitro*** studies in experimental biology are those that are conducted using components of an organism that have been isolated from their usual biological surroundings in order to permit a more detailed or more convenient analysis than can be done with whole organisms.]

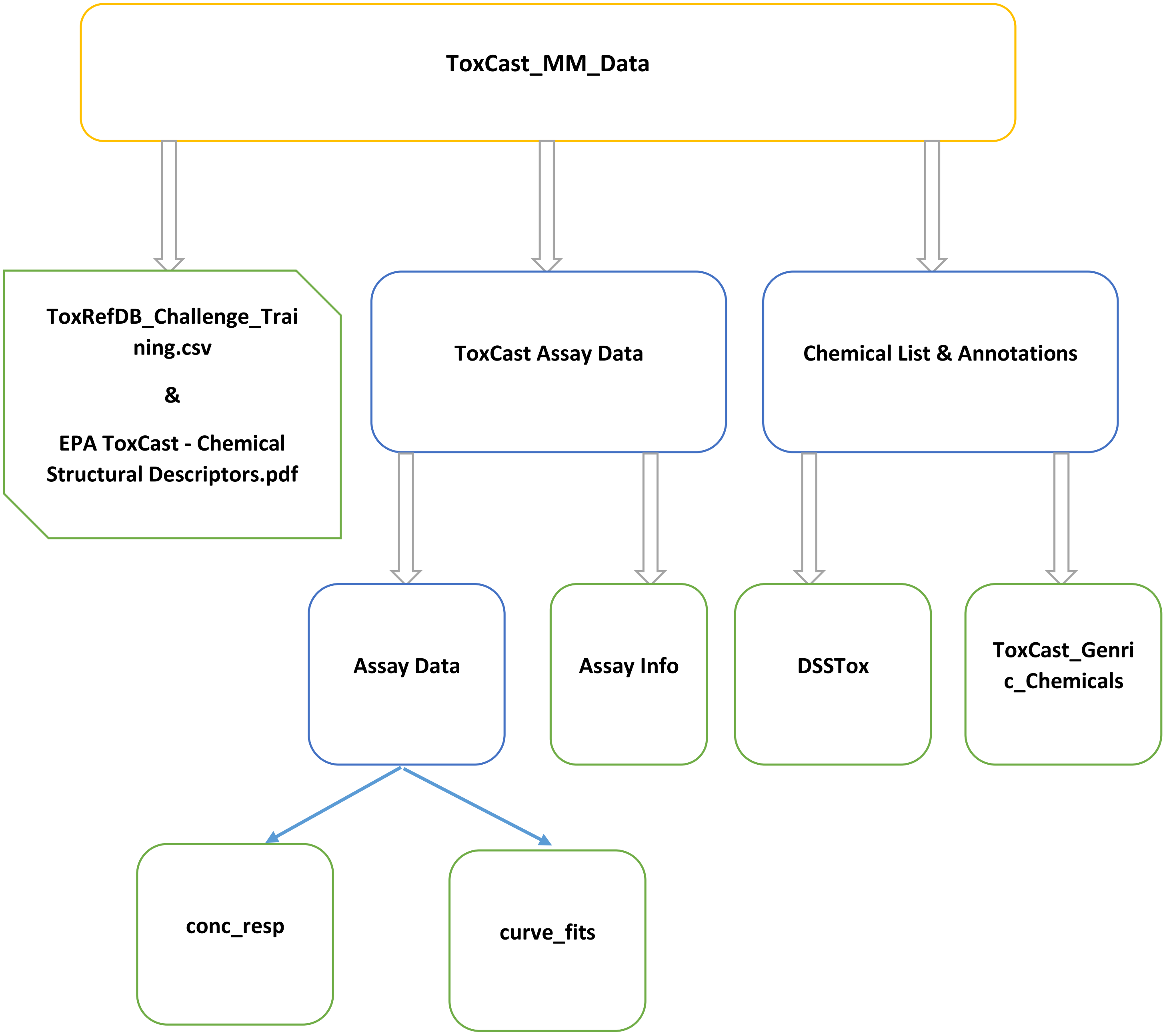
The EPA procured ~3600 unique chemicals, of which ~1000 have been tested in the complete set of ToxCast Phase II assays (and some additional assays in Phase I that were subsequently discontinued), and a further ~880 in an endocrine-related subset of the ToxCast Phase II assays; note that we refer to this ~3600 EPA inventory, henceforth, as “ToxCast”. Some of the files we provide will contain ~1000 chemicals (the main ToxCast Phase I and II data) and some will contain ~1800 chemicals (ToxCast Phases I and II augmented with the “E1K” set). In most cases, because we have tested multiple samples of the same chemical, there will be additional data points. An important part of the ToxCast project is linking in vitro bioactivity with whole animal toxicity results for a common set of chemicals. The ToxRef project has compiled guideline-level in vivo toxicity data into ToxRefDB.

“E1K” in any file name indicates that the file contains ToxCast Phase I, II and E1K chemicals.

2.0 Data Organization and Understanding

2.1 Directory Chart

The data is organized into different directories in the following manner:


ToxCast_MM_Data
ToxRefDB_Challenge_Training.csv
&
EPA ToxCast - Chemical Structural Descriptors.pdf
ToxCast Assay Data
Chemical List & Annotations
Assay Data
Assay Info
DSSTox
ToxCast_Genric_Chemicals
conc_resp
curve_fits

- All the green boxes above indicate either the file or the last directory in the hierarchy. Each last directory contains a sub- folder named “**csv**” inside which all relevant files are available.

2.1.1 ToxRefDB_Challenge_Training.csv

This is the training file that will be used to train the model. This file contains data from ToxRefDB which is obtained from animal toxicity studies on chemicals. This file contains minimal information that will help you map all the other features from various available data sources. Following are the field in this file:

DSSTox_GSID: Generic Substance ID from DSSTox – this is generally 1:1 with a CASRN.

CARSN: Chemical Abstracts Registry Number; in cases where a CASRN is unavailable for a chemical, we assign a CASRN-like value, e.g., “NOCAS_20182”, where 20182 is the DSSTox_GSID value.

chemical_name: Name of chemical in question.

systemic_adjusted_negative_log_lel: This is the field whose values are to be **predicted** in this challenge. It represents the Lowest Effective dose Level (LEL) of chemical.

Please Note: For all the references to chemicals across the files, it is recommended to use CARSN ID only. CARSN stands for Chemical Abstracts Registry Number. There are different types of ids available but as the training set contains only CARSN, using that id across all files as a key for mapping various features is the suggested approach. CARSN is available in most of the files except a couple – handling it will be explained below.

EPA ToxCast - Chemical Structural Descriptors.pdf

This file contains a lot of resources and information about many different libraries and tools that can be used for generating chemical descriptors and fetch chemical structures for different chemical compounds. The members are encouraged to use any of these libraries as deemed necessary to incorporate chemical structure into prediction models.

Using chemical structures as features in the model and thereby providing structural knowledge to the algorithm is expected to aid the prediction as chemical structure of chemical is expected to be linked with the LEL. The two main sections of importance in this file are: 1.) Libraries and Tools and 2.) Databases. Please ask in forum if anything is unclear about this document. This document was obtained thorough an idea generation contest held by TopCoder recently and we have full support available for it.

ASSAY DATA

For each assay-chemical combination, the data was fit to a Hill curve to yield estimates of the potency (the AC50 or concentration at 50% activity) and efficacy (Emax or maximum response).

2.1.2 ToxCast Assay Data/Assay Info/csv

This directory contains description about study design of various assays and their targets. It covers information like assay sources, the organism targeted in assays, the tissue that was targeted in particular assay, assay format types, etc. These files can be considered as metadata of various assays and contains a lot of text data making it quite less relevant to the prediction challenge. But this file can be used in the later phase to derive analysis of biological evidence to the prediction results.

2.1.3 ToxCast Assay Data/Assay Data/Curve fits/Csv

This folder is again divided into three different sub-folders and an AC50 Summary file is provided.

A.) **ToxCast_Summary_AC50_2013_12_10_NO_BSK.csv:** This file provide a summary of active/inactive calls for the entire dataset. The file provides this high level summary for the entire 1800 chemical ToxCast data set – the cells with value 1.00E+06 are inactive, the cells with other smaller values are AC50 – potent (low) values while NA are not applicable. In the original file, there is a color-coding with cells color-coded by AC50 – potent (low) values in green and inactive in white (with value "##"). But csv conversion removed all the formatting and changed the "##" value to 1.00E+06.

Please Note: All the valid values in summary files are in **micro molar** units.

B.) **Assay Fit Directory:** These files contain one row per assay-chemical-sample combination and have extensive information on the curve fitting parameters, confidence intervals and QC metrics. The fields in these files are described in the appendix. The assay fit files are provided separately for each the assays done by each of the following organizations: ACEA, Apredica, Attagene, NCGC, NovaScreen and Odyssey.

C.) **Pathway Related Assays:** These files provide following values for several pathway related assays:

AC50 values and AC50 mod values,

Emax values,

Level 7 and level 8 values (these levels correspond to the levels in data processing pipeline developed by EPA [1].),

Max Concentration values

Top (t) and Slope (W) of the Hill Curve Values.

The columns in these files are the pathways and can be considered as black box features.

D.) **Result Matrix Files**: These files provide the following result values for different assays:

AC50 values and AC50 mod values,

Emax values,

Level 7 and level 8 values (these levels correspond to the levels in data processing pipeline developed by EPA [1].),

Max Concentration values

Top (t) and Slope (W) of the Hill Curve Values.

2.1.4 ToxCast Assay Data/Assay Data/Conc Resp/Csv

In this folder, there is a separate .csv file for every assay used to screen the chemicals. Each .csv file contains data for one assay for all chemicals. For each chemical and assay pair, there are multiple concentrations of chemical tested. So for each chemical and assay pair, you will have multiple concentrations with multiple responses (thus multiple rows belonging to same chemical). This is a very detailed data and a very huge data too. This data can be used for trying various feature combinations from different assay responses to build complex models.

The columns of major interest in all these files are CARSN, conc and response. These files have large amount of data samples and very few features.

Chemical List and Annotations

2.1.5 ToxCast Assay Data/Chemical List & Annotations/ToxCast_Generic_Chemicals/csv

This file provides information about the chemicals at generic level. The columns of interest are CARSN, name, UseCategory, molecular weight and chemical formula which may be a contributing factor in LEL predications.

2.1.6 ToxCast Assay Data/Chemical List & Annotations/DSSTox/csv

A.) **toxprint_v2_vs_TOX21S_v4a_8599_03Dec2013.csv:** This file provides information about the presence or absence of different kinds of chemical bonds in the atomic structure of the chemical. The only ID available in this file is DSSTox_CID which means this file will require reference to general file for mapping chemical name.

B.) **TOX21S_v4a_8599_11Dec2013.csv:** These files contain the most recent publication of all chemicals in ToxCast library and contains the complete structure information in nomenclature like IUPAC and SMILES structures along with chemical names. The column names are self-explanatory.

3.0 Useful Information

The above provided details will provide an easy understanding of the available features. But if you are interested to look at the detailed explanation of the whole data processing pipeline, please refer to [1].

At this point, we ran an internal analysis of the usability of this data and we have found that there are several important files which contributes towards LEL prediction. We are using those files as input to this match and they are the following:

1.) ToxRefDB_Challenge_Training.csv

2.) All the Result Matrix files (Section 2.1.3 D)

3.) TOX21S_v4a_8599_11Dec2013.csv (structure file)

4.) ToxCast_Generic_Chemicals_2013_12_10.csv (general chemical information)

5.) ToxCast_Summary_AC50_2013_12_10_NO_BSK.csv (AC50 summary file).

You are encouraged to use the remaining data as additional features and try various combinations to build complex features to improve the accuracy of the prediction.

4.0 Appendix

There are many columns with different names and field values across the above mentioned files. Here, we provide a consolidated list of all the fields and interpretation of their values as you may come across any of them.

Level Files (Many files in assay data under curve fit directory):

- sample_tech_rep_id = Index of the sample_rep_id (see below) based on the number of technical replicates for that assay; General assay reproducibility statistics performed at this level (e.g., correlation among technical replicates)
- sample_rep_id = Individual sample replicate ID; This is the blinded ID sent to assay contractors. Sample replicates come from the same stock solution and lot/batch and were intended additions to the sample-chemical library; Replicate analysis (from same lot/batch) performed at this level (e.g., hit-call concordance).
- sample_id = Base sample ID which equates to a sample vial from one stock solution and lot/batch; Primary analysis performed at this level.
- chemical_id = DSSTox GSID; Additional reproducibility performed at this level to evaluate impact of separately sourced chemicals.
- chemical_casrn = CAS registration number.
- chemical_name = Chemical name; This may be a trade name, generic name, common name, IUPAC name, or some combination thereof.
- assay_name = Unique assay component endpoint name (breaks out by readout, timepoint, curve fit direction, etc).
- AC50.mod = AC50 modifier used to distinguish hit confidence at a terse level;
- “==”-met basic requirements of Hill model with some minimal confidence in T and B
- “><”-Not a hit and no model fit
- “<=”-AC50 set to low concentration tested as no model could be determined but high levels of activity were observed
- “>=”-Hill model was fit but with little confidence in T (top) of curve as the model had a Hill slope of 8 and had only a single point determining the hit call
- “<>”-Significant activity was observed but the data was noisy or the Hill model did not meet certain criteria. This assumes the AC50 should be somewhere in the tested concentration range but there is little confidence in the point estimate
- B = Bottom (minimum asymptote) of curve (parameter of Hill model)
- T = Top (maximum asymptote) of curve (parameter of Hill model)
- logAC50 = 50% maximal response concentration (log uM) (parameter of Hill model)
- AC50 = Converted 50% maximal concentration from Hill model (uM)

- W = Hill slope (parameter of Hill model)
- Emax = Maximal average response at a given concentration (outlier removed)
- T2B = T-B (Top of curve minus Bottom of Curve)
- Rsq = R-squared of the Hill model
- B.std.error = Standard Error of the B parameter (from summary(nls))
- T.std.error = Standard Error of the T parameter (from summary(nls))
- logAC50.std.error = Standard Error of the logAC50 parameter (from summary(nls))
- W.std.error = Standard Error of the W parameter (from summary(nls))
- B.pval = P-value (F-statistic) of the B parameter (from summary(nls))
- T.pval = P-value (F-statistic) of the T parameter (from summary(nls))
- logAC50.pval = P-value(F-statistic) of the logAC50 parameter (from summary(nls))
- W.pval = P-value (F-statistic) of the W parameter (from summary(nls))
- logAC50.confint.lower = Lower 95%-tile confidence interval around logAC50 parameter (confint(nls[“logAC50”]))
- logAC50.confint.upper = Upper 95%-tile confidence interval around logAC50 parameter (confint(nls[“logAC50”]))
- AC50.confint.lower = Transformed value from logAC50
- AC50.confint.upper = Transformed value from logAC50
- AC50.2se.lwr = AC50 minus two times the transformed logAC50 standard error
- AC50.2se.upr = AC50 plus two times the transformed logAC50 standard error
- B.confint.lower = Lower 95%-tile confidence interval around B parameter (confint(nls[“B”]))
- B.confint.upper = Upper 95%-tile confidence interval around B parameter (confint(nls[“B”]))
- T.confint.lower = Lower 95%-tile confidence interval around T parameter (confint(nls[“T”]))
- T.confint.upper = Upper 95%-tile confidence interval around T parameter (confint(nls[“T”]))
- W.confint.lower = Lower 95%-tile confidence interval around W parameter (confint(nls[“W”]))
- W.confint.upper = Upper 95%-tile confidence interval around W parameter (confint(nls[“W”]))
- Min.conc and max.conc = minimum and maximum tested concentration with a usable (not flagged) data point
- Min.resp and max.resp = minimum and maximum response of all non-flagged responses (max.resp does not equal Emax as Emax is the max average response at a concentration and not the single maximal response)
- Lec.resp.cutoff = the resp value required to determine an LEC and a minimal requirement to be considered a hit
- Ac50.resp.cutoff = the resp value required to determine an AC50 and a minimal requirement to be considered a hit (e.g., 10 times the baseline MAD); these values generally range from 15-50% and vary primarily due to the background noise level of the assay.
- Rsq.cutoff = Secondary hit-calling cutoff which requires the Hill model to have an r-squared value above this cutoff to determine an AC50 and have an AC50.mod equal to “==”.
- LEC = Lowest Effective Concentration, First concentration in which the average response at a concentration is above the lec.resp.cutoff.
- AC10-90 = ACXX <<- function(T2B,AC50,W,xx){XX <- xx/100;F <- XX*(T2B);ACXX <- ((F/(T2B-F))^(1/W))*AC50; return(s3(ACXX))}AC20
- y50 = Efficacy at the AC50 [(0.5 * T2B)+B]

- line.x = Hill model line X (concentration) coordinates, Captured to assist in plotting curve if equation is not available
- line.y = Hill model line Y (efficacy) coordinates, Captured to assist in plotting curve if equation is not available
- converged = Did the Hill model converge? TRUE/FALSE
- convergence.message = Convergence message from NLS
- line.y.pi.lwr & upr = Hill model line of the prediction interval (below and above the Hill model)
- AC50.pi.lwr = Lower prediction interval around AC50
- AC50.pi.upr = Upper prediction interval around AC50
- RSQUARED_TEST = TRUE or FALSE; Did the Hill model meet the R-Squared criteria (>0.5)
- T2B_TEST = TRUE or FALSE; Did the Hill model meet the T-B test based on the response cutoff (ac50.resp.cutoff)
- Emax_TEST = TRUE or FALSE; Did the max response meet the response cutoff (ac50.resp.cutoff)
- OVERALL_TEST = TRUE or FALSE; Did the model meet all 3 criteria or a special case (such as the "<>" AC50.mod case)
- Level6.ac50.hitcall = TRUE or FALSE; Did the model meet all 3 criteria or a special case (may override special case from TRUE to FALSE)
- Level6.lec.hitcall = TRUE or FALSE; Did the Emax rise above the lec.resp.cutoff
- Level7.ac50.hitcall = TRUE or FALSE; Based on external data or confounding properties of the assay level6.ac50.hitcall can be changed systematically
- Level7.lec.hitcall = TRUE or FALSE; Based on external data or confounding properties of the assay level6.lec.hitcall can be changed systematically
- Level7.ac50.notes = Provides reason for any change
- Level7.lec.notes = Provides reason for any change
- Level8.ac50.hitcall = TRUE or FALSE; Based on manual review, hit calls can be changed; justification required
- Level8.lec.hitcall = TRUE or FALSE; Based on manual review, hit calls can be changed; justification required
- Level8.ac50.notes = Provides reason for any change
- Level8.lec.notes = Provides reason for any change

Assay Curve Fit Files Columns

- assay_fit_id – database id for this data point
- assay_id – database id for the assay
- datafile_id – database id for the input data file
- substance_id – database ID for the substance / sample
- assay_name – source_name_aid
- level – 6,7,8
- sample_id – sample ID traceable to the bottle, supplier/lot/batch and stock solution
- sample_rep_id – sample ID traceable to the bottle with replicate digit
- chemical_id – DSSTox GSID
- chemical_casrn - CASRN

- chemical_name – preferred chemical name
- AC50 – AC50
- AC50_mod, AC50_mod_1 – modifier (==,<=,>=,<>,><)
- AC50_confint_lower, AC50_confint_upper – AC50 confidence intervals
- AC50_2se_lwr, AC50_2se_upr - +/- 2 standard error for AC50
- AC50_pi_lwr, AC50_pi_upr - ???
- logAC50 – log10(AC50)
- logAC50_std_error – standard error for the logAC50
- logAC50_pval – p-value for the logAC50
- logAC50_confint_lower, logAC50_confint_upper – confidence intervals around the logAC50
- B, T, W – remaining Hill-fit parameters (Bottom, Top, Slope)
- B_std_error, B_pval, B_confint_lower, B_confint_upper, T_std_error, T_pval, T_confint_lower, T_confint_upper, W_std_error, W_pval, W_confint_lower, W_confint_upper – statistical parameters around B,T, W
- Emax – maximum mean response across concentrations
- T2B – T-B
- Rsq – R-squared for fit of data to the Hill-curve
- LEC – Lowest Effect Concentration – lowest tested concentration at which the mean response is above the response cutoff.
- AC10, AC20, AC30, AC40, AC60, AC70, AC80, AC90 – interpolate ACx values for the Hill-curve, corresponding to response at x% of T
- y50 – actual response of Hill curve at 50% of T
- rAC50 – relative AC50 – concentration at which the response is 50% absolute rather than 50% of T
- line_x, line_y – x and y vectors to draw the Hill curve
- line_y_pi_lwr, line_y_pi_upr - ???
- min_conc, max_conc – min and max concentrations tested
- min_resp, max_resp – min and max values of the response for any concentration, any technical replicate
- converged – Boolean, did the fitting algorithm converge?
- convergence_message – any error message regarding convergence
- resp_cutoff – response cutoff – to be a Level 6 hit, one of the mean responses has to be greater than this value. Typically 10 MAD around the baseline noise
- rsq_cutoff – cutoff on R-squared for declaring that a good Hill-curve fit is achieved
- T2B_TEST – test on Top-to-Bottom (Boolean)
- RSQUARED_TEST – teston R-squared (Boolean)
- Emax_TEST – test on Emax (Boolean)
- OVERALL_TEST – Overall fitting test (Boolean)
- conc_bag - ???
- rm_conc_denominator - ???
- rm_hit_perc - ???
- HIT_CALL_LEVEL6, HIT_CALL_LEVEL7,HIT_CALL_LEVEL8 – Boolean (deprecated and replaced by parameters below)

- ind_avg – Was the fitting done at the individual technical replicate level or averaging over technical replicates
- status – status in the database
- ac50_resp_cutoff, lec_resp_cutoff – extra cutoff parameters
- level6_ac50_hitcall, level6_lec_hitcall, level7_ac50_hitcall, level7_lec_hitcall, level8_ac50_hitcall, level8_lec_hitcall – Hitcall (TRUE/FALSE)
- chemical_set
- final_hit_call, final_hit_call_note, level7_ac50_notes, level7_lec_notes, level8_ac50_notes, level8_lec_notes – Notes on override calls
- promiscuity, viability_promiscuity, bla_promiscuity, luc_promiscuity – special flags for NCGC assays

Special Values:

For post-processing code, it is useful to have all data values be numbers, so special cases (missing data, etc.) are represented by special values as follows:

- $TOXCAST_PHASE_1_INACTIVE=1000000 – this is the default inactive value for any Phase I data that has not gone through the current workflow
- $NVS_PREFILL=1000201 – because Novascreen is only run in concentration-response for single-point actives, we fill in the rest of the chemical-by-assay matrix with this value to indicate that single point was run but failed. No corresponding values are in the workflow data files
- $VALUE_NA=1000901 – the original data files had a value of “NA”
- $VALUE_NAN=1000902 – the original data files had a value of “NaN”
- $VALUE_INF=1000903 – the original data files had a value of “Inf”
- Any other value 1000000+ is inactive
- 2000000 – missing data

5.0 Reference

[1] ToxCast Release Document, December 2013. [To be provided in forums]